DOCUMENT RESUME

ED 110 524 TM 004 851

TITLE	Proceedings of the Annual Conference of the Military Testing Association (16th).
PUB DATE	[Oct 74]
NOTE	663p.; Papers presented at the Annual Meeting of the Military Testing Association (16th, Oklahoma City, Oklahoma, October 21-25, 1974)
EDRS PRICE	MF-$1.08 HC-$33.64 PLUS POSTAGE
DESCRIPTORS	Computers; Interest Tests; Measurement Techniques; *Military Personnel; Military Science; *Military Training; Performance Tests; *Personnel Selection; Program Effectiveness; Statistical Analysis; Task Analysis; Task Performance; Test Construction; Testing; *Testing Problems; Test Results; *Test Validity

ABSTRACT

The military utilizes tests extensively and should be aware of current testing problems. A conference composed of members from the business, educational, and military communities, both foreign and domestic, was held and these major issues addressed: (1) The Development of Training Requirements; (2) Personnel Assignment Using Interest Inventories; (3) Flight Simulators--Their Development and Value in Training; (4) Test Development; (5) New Test Types and Test Validation; (6) Performance Measurement and Statistical Methodologies; (7) Training Methodologies; (8) Occupational Analysis in the Military Establishments; (9) Using Test Results in the Training Process; (10) Computer Assisted Training and Testing; (11) Personnel Assessment; (12) Management and Measurement Problems; (13) The Care and Feeding of Test Constructors; and (14) Evaluating Job Proficiency. Issues of public policy, vocational advancement, peer and preceptor ratings, student evaluations, problem identification, classification procedures, and licensing innovations were included in the general sessions. (BJG)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED110524

PROCEEDINGS

16th Annual Conference

of the

Military Testing Association

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

Coordinated By:

U. S. COAST GUARD INSTITUTE
P. O. Substation 18
Oklahoma City, Oklahoma 73169

HILTON INN WEST
Oklahoma City, Oklahoma
21 October - 25 October 1974

TM 004 851

2/3

FOREWORD

The papers presented at the Sixteenth Annual Conference of the Military Testing Association came from the business, educational, and military communities, both foreign and domestic. The papers reflect the opinions of their authors only and are not to be construed as the official policy of any institution, government, or branch of armed service.

TABLE OF CONTENTS

Page

Official Program. 1

Keynote Address . . . . RADM Robert W. Durfey 12

SECTION I: The Development of Training Requirements. 16

"Development and Application of an Algorithm for Translating Job Information into Technical Training Requirements". . . . . .MAJ W.S. Sellman 17
SQ LDR J.W.K. Fugill

"One of the Missing Links. . . . . .CDR B. Cormack 27

"Structural Task Analysis--The Bridge Between Selection and Training". . . . . Mr. J.W. Dees 31

SECTION II: Personnel Assignment Using Interest Inventories 45

"Interpretive Report for the Strong-Campbell Interest Inventory" . . . . .Dr. C.B. Johansson 46

"Clemans' Lambda-Based Scales of the Navy Vocational Interest Inventory". . . . . Dr. J.E. Dann 100

"Development and Evaluation of Strong Vocational Interest Blank Scales for Academic Major Areas at the U.S. Naval Academy". . . . . .Mr. W.A. Sands 112
LCDR R.D. McCullah

"Interest Measurement in Army Classification". . . . . Dr. M.A. Fischl 127

SECTION III: Flight Simulators--Their Development and Value in Training 134

"Simulator Motion as a Factor in Flight Simulator Training Effectiveness" . . . . Mr. R.S. Jacobs 135

"Development of an RPV Operator Trainer" . . . . Mr. L.E. Ehrhardt 149

Page

"Performance Evaluation: The Use of Scoring Systems in Adaptive Training" . . . . CAPT O.H. Deberg, SQ LDR C.J. Hyatt 155

SECTION IV: Test Development 177

"Use of Occupational Survey Data in Developing Outlines for Specialty Knowledge Tests" . . . . Ms. N. Lewis 178

"The Effect of Format Change in Item Difficulty" . . . . Mr. J.L. Boyd 184

"A 1" x 3½" Keyhole on the World of MOS Testing" . . . . Mr. A.G. Hermansen 187

SECTION V: New Test Types and Test Validation 193

"A New Sequential Multiple Choice Skill Related Testing Device" . . . . Mr. F.B. Braun, Mr. J.E. Tindall, Jr. 194

"Behavior Toxicology: A New Frontier for Behavioral Testing" . . . . Dr. T.C. Tuttle 212

"Entrance Scores and Training Performance for Two Groups of Coast Guard Input" . . . . Dr. A.S. West, Prof. R.E. Williams 225

"Selection of Students for the Naval Postgraduate School" Dr. R.S. Elster, Dr. W.H. Githens 235

SECTION VI: Performance Measurement and Statistical Methodologies 242

"Automated Performance Measurement Used in Quality Control of the U.S. Coast Guard Aviator Training System" . . . . Mr. H.K. Povenmire 243

"An Automated Air Crew Station Anthropometric Unit" . . . . Ms. C. Bohn 256

"Prediction from Contingency Tables Using Joint Liklihoods" . . . . LCDR W.S. Shields 261

Page

SECTION VII: Training Methodologies. 273

"Boating Safety Education: Three Routes to a Common Goal" . . . . .Ms. S.C. Irish 274
BMCM J.L. Wilson

"Student Response in Television Instruction" . . . . Mr. F.W. Brashear 280

SECTION VIII: General Session 288

"Public Policy Concerning the Use of Tests". . . . . .Dr. R.L. Lewiski 289

"The Common Bond in Peer and Preceptor Ratings" . . . . CAPT J.C. Helmick 314

"The Vocational/Advancement Scheme of the Federal Armed Forces Association". .LT COL J. Fricke 335

SECTION IX: Symposium: Occupational Analysis in the Military Establishments 339

"Occupational Analysis in the United States Air Force" . . . . Dr. W.E. Driskill 340

"Naval Occupational Task Analysis Program" . . . . Mr. M.D. Callahan 346

"Occupational Analysis in the Royal Navy". . . . . CDR C.J. Nicoll 357

"The Royal Australian Air Force Occupational Measurement Program". . . . . .MAJ W.S. Sellman 362
SQ LDR J.W.K. Fugill

SECTION X: Using Test Results in the Training Process. 371

"Test Results: The Key to the Management of Training". . . . . Mr. R.N. Johnson 372

"Task Analysis Based Testing as a Measure of Training Effectiveness" . . . . LCDR C.L. Walker 381

7

Page

"The Interface of Testing and Training: Air Force Experience 1972-1974" . . . . .COL L.D. Kaapke 386

SECTION XI: Computer Assisted Training and Testing 392

"Contributions to Objective Measurement and Evaluation of Trainee Competency" . . . . .Dr. W.J. Moonan 393

"The Use of Tactical Computers to Provide Weapons and Tactics Training to Combat Infantry NCO's: Results of a Field Test" . . . . .Dr. W.G. Hoyt 425
Mr. J.D. Baker

SECTION XII: Personnel Assessment 441

"Assessment and Counseling for Leadership Development" . . . . .Dr. J.A. Caviness 442

"Dimensional Analysis of Leadership Measures from a Training Program" . . . . .Dr. R.G. Downey 452

"Assessment of Overseas Performance". . . . . .Mr. T. Yellen 459

SECTION XIII: Management and Measurement Problems. 463

"Response Bias in Worker Task Inventories". . . . . MAJ W.E. Slack 464

"The Conflicting Demands of Go-No-Go Performance Criteria and the Requirements for Rank Ordering Students". . . . . . .Dr. W. Scanland 496

SECTION XIV: Symposium: The Care and Feeding of Test Constructors. 501

"The Perspective of a Testing Psychologist Who Conducts Orientation Programs for New Test Constructors" . . . . .Mr. W. Birdsall 502

"The Perspective of a Military Supervisor of Test Constructors Who Are Temporarily Brought in from the Field to Write Tests" . . . . 1LT D. Vaughan 506

"The Perspective of a Civilian Test Psychologist Working with Test Constructors" . . . . .Mr. C. Smith 510

"The Perspective of a Non-Commissioned Officer Who is Actually Constructing Tests" . . . . .ADCM J.R. Harris 514

"Results of a Service-Wide Survey of Enlisted Test Constructors" . . . . LT COL G. Oetting 518

SECTION XV: General Session 539

"The Development of a Student Evaluation of Course and Instructor Report". . . . . LT COL G.J. Carpenter 540

"Problem Checklists, A Pragmatic Approach for Fostering Unit Effectiveness" .Dr. F.W. Timmerman 552

"Classification of Army Officer Duty Positions: Implications for Training Requirements" . . . . .Dr. A.C.F. Gilbert 559

"U.S. Merchant Marine Licensing Examinations: The Modernization of a Traditional Mission" . . . . .Mr. R.C. Willing 576
LT D. McDevitt

SECTION XVI: Symposium: Evaluating Job Proficiency. 581

"Functionally Integrated Performance Testing" . . . . Mr. W.C. Osborn 582
Mr. J.H. Harris
Mr. J.P. Ford

"Test Fidelity--Congruence of Test Tasks with Real-World Performance Tests. . . . . . .Dr. A. Mirabella 590
Mr. E. Johnson

"Development of Criterion Ratings to Measure Job Performance". . . . . Mr. C.P. Hahn 596

"Scaling Techniques to Reduce Rater Leniency on Navy Enlisted Performance Evaluation Forms". . . . . .Ms. M.H. Royle 616
Mr. D.W. Robertson

Steering Committee Meeting Minutes. 632

9

Page

Business Meeting Minutes. 634

Military Testing Association By-Laws. 635

List of Conferees 640

Index 648

16th Annual Conference of the MILITARY TESTING ASSOCIATION

Officers, Chairmen, and Committees

MTA President
COMMANDER KENNETH M. ROUGHGARDEN

MTA Secretary
MR. JOHN A. BURT

Chairmen:	Committee:
CWO4 EDWARD L. AMOS	Financial
CWO3 GALE E. ZABLE	Administration
MR. RICHARD C. WILLING	Program
CWO3 JOHN A. HALDEMAN	Visual/Audio Aids
LT DAVID E. MCDEVITT	Social
LTJG JAMES D. JONES	Publicity

16th Annual Conference
MILITARY TESTING ASSOCIATION

21-25 October

MONDAY, OCTOBER 21

Lobby
0900-1800 Registration

Presidential Suite
(Room 404)
1600-2000 Steering Committee Meeting

TUESDAY MORNING, OCTOBER 22

Palomino Room
0900-1000 Conference called to order

0900-0915 Greetings by Coast Guard Institute
Commanding Officer
CDR KENNETH M. ROUGHGARDEN, USCG

0915-1000 Keynote Address
RADM ROBERT W. DURFEY, USCG
Chief, Office of Personnel
U.S. Coast Guard

Appaloosa Room
1000-1030 Coffee

Palomino Room
1030-1200

1030-1040 MTA Research Project
LT COL GEORGE D. OETTING, USAF

1040-1145 The Development of Training Requirements
Moderator:
LCDR FREDERICK J. GRADY, USCG

"*Development and Application of an Algorithm for Translating Job Information into Technical Training Requirements*" (25 min.)
MAJ WAYNE S. SELLMAN, USAF, and
SQ LDR JOHN W.K. FUGILL, AUSTRALIAN ROYAL AIR FORCE
Presented By: SQ LDR W.J. WATSON, RAAF

"*One of the Missing Links*" (20 min.)
CDR BRUCE CORMACK, CANADIAN FORCES

Tuesday Morning

Palomino Room — *"Structural Task Analysis--The Bridge Between Selection and Training"*
(20 min.)
JAMES W. DEES, HumRRO

1145-1200 — Announcements

1200-1315 — Lunch

TUESDAY AFTERNOON, OCTOBER 22

Quarterhorse Room

1315-1445 — Personnel Assignment Using Interest Inventories
Moderator:
LT DAVID MCDEVITT, USCG

"Interpretive Report for the Strong-Campbell Interest Inventory" (20 min.)
DR. CHARLES B. JOHANSSON
National Computer Systems

"Clemans' Lambda-Based Scales of the Navy Vocational Interest Inventory"
(15 min.)
JOYCE E. DANN, NPRDC

"Development and Evaluation of Strong Vocational Interest Blank Scales for Academic Major Areas at the U. S. Naval Academy" (20 min.)
WILLIAM A. SANDS, NPRDC and
LCDR ROBERT D. MCCULLAH, USN, Ph.D
Presented by: LCDR McCullah

"Interest Measurement in Army Classification" (25 min.)
DR. M. A. FISCHL, USA Research
Institute, Arlington VA

1500-1635 — Flight Simulators--Their Development and Value in Training
Moderator:
LT THOMAS PEARSON, USCG

"Simulator Motion as a Factor in Flight Simulator Training Effectiveness"
(30 min.)
ROBERT S. JACOBS, Aviation Research
Lab, Univ. of Illinois

"Development of an RPV Operator Trainer"
(20 min.)
LOUIS E. EHRHARDT, U.S. Naval Missile
Center

Tuesday Afternoon

Quarterhorse Room

"*Performance Evaluation: The Use of Scoring Systems in Adaptive Training*" (45 min.)
CAPT OAK H. DEBERG, USAF and
SQ LDR CHRISTOPHER J. HYATT, RAF

Palomino Room
1315-1445

Test Development
Moderator:
LT TIM FLANAGAN, USCG

"*Use of Occupational Survey Data in Developing Outlines for Specialty Knowledge Tests*" (20 min.)
NANCY A. LEWIS, USAF Occupational Measurement Center

"*The Effect of Format Change in Item Difficulty*" (15 min.)
JOSEPH L. BOYD, JR., Educational Testing Service

"*A 1" x 3¼" Keyhole on the World of MOS Testing*" (25 min.)
ARTHUR G. HERMANSEN, USAEEC
Presented by: Clay V. Brittain, USAEEC

1500-1630

New Test Types and Test Validation
Moderator:
LTJG ROBERT P. MOORE, USCG

"*A New Sequential Multiple Choice Skill Related Testing Device*" (25 min.)
F. B. BRAUN and J. E. TINDALL, JR., Data-Design Laboratories

"*Behavior Toxicology: A New Frontier for Behavioral Testing*" (15 min.)
DR. THOMAS C. TUTTLE
Westinghouse Behavioral/Safety Center

"*Entrance Scores and Training Performance for Two Groups of Coast Guard Input*" (25 min.)
DR. ANITA S. WEST and
RALPH E. WILLIAMS, University of Denver Research Institute

"*Selection of Students for the Naval Postgraduate School*" (25 min.)
DR. R. S. ELSTER and
DR. W. H. GITHENS, U.S. Naval Postgraduate School (NPRDC)

Tuesday Afternoon

Arabian Room
1315-1445 Performance Measurement and Statistical Methodologies
Moderator:
LCDR ELDON L. BEAVERS, USCG

"Automated Performance Measurement Used in Quality Control of the U.S. Coast Guard Aviator Training System" (25 min.)
H. KINGSLEY POVENMIRE, USCG Aviation Training Center

"An Automated Air Crew Station Anthropometric Unit" (20 min.)
LCDR HARV GREGOIRE, USN and
CAROLE BOHN, Texas Technical University
Presented by: Ms. Bohn

"Prediction from Contingency Tables Using Joint Liklihoods" (25 min.)
LCDR W. S. SHIELDS, Royal Military College of Canada

1500-1630 Training Methodologies
Moderator:
LTJG PHILLIP CHASE, USCG

"Boating Safety Education: Three Routes to a Common Goal" (30 min.)
BMCM JACK L. WILSON, USCG, and
SALLY C. IRISH, USCG Boating Safety
Presented by: Ms. Irish

(Not Presented) *"Training Requirements and Control in Manpower Structure--Studies of the Armed Forces"* (35 min.)
COL ALFRED BARKHOFF, German Armed Forces

"Student Response in Television Instruction" (25 min.)
FENTON W. BRASHEAR, USA Ordnance Center and School

1900-2100 Gazebo Room - Social Hour

WEDNESDAY MORNING, OCTOBER 23

Palomino Room
0830-0930 General Session
Moderator:
LT WILLIAM G. WOHLFARTH

Wednesday Morning

Palomino Room	"*Public Policy Concerning the Use of Tests*" (15 min.) DR. RAYMOND L. LEWISKI, Extension Course Institute, Gunter AFS, AL
	"*The Common Bond in Peer and Preceptor Ratings*" (20 min.) CAPT JOHN C. HELMICK, USA Infantry School, Ft. Benning, GA Presented By: Mr. James Gerber
	"*The Vocational/Advancement Scheme of the Federal Armed Forces Association*" (20 min.) LT COL J. FRICKE, Federal Armed Forces West Germany
Appaloosa Room 0930-1000	Coffee
Palomino Room 1000-1200	Symposium: Occupational Analysis in the Military Establishments Chairman: DR. WALTER E. DRISKILL, USAF Occupational Measurement Center Participants: COL JAMES R. GALLMAN, U.S. Marine Corps; B. MICHAEL BERGER, U.S. Army; MAURICE D. CALLAHAN, U.S. Navy; CDR BRUCE CORMACK, Canadian Forces; RICHARD S. LANTERMAN, U.S. Coast Guard; CDR COLIN J. NICOLL, British Royal Navy; MAJ WAYNE S. SELLMAN, U.S. Air Force for Australia
1155-1200	Announcements
1200-1315	Lunch

WEDNESDAY AFTERNOON, OCTOBER 23

Quarterhorse Room 1315-1500	Using Test Results in the Training Process Moderator: LCDR MARVIN E. WILMOTH, USCG
	"*Test Results: The Key to the Management of Training*" (30 min.) ROBERT N. JOHNSON, U.S. Army Institute of Administration, Ft. Ben Harrison, IN

Wednesday Afternoon

Quarterhorse Room

(Not Presented)

"Army NCO Qualification Test Procedures as Part of the Sergeant Training Program" (25 min.)
LT COL H. E. SEUBERLICH, Federal Armed Forces Association, West Germany

"Task Analysis Based Testing as a Measure of Training Effectiveness" (20 min.)
LCDR C. C. WALKER, USN

"The Interface of Testing and Training: Air Force Experience 1972-1974" (25 min.)
COL LYLE D. KAAPKE, USAF Occupational Measurement Center

Palomino Room
1315-1500

Computer Assisted Training and Testing
Moderator:
LT HARRY L. IMMER, USCG

"Contributions to Objective Measurement and Evaluation of Trainee Competency" (25 min.)
DR. WILLIAM J. MOONAN, USN (NPRDC)

"The Use of Tactical Computers to Provide Weapons and Tactics Training to Combat Infantry NCO's: Results of a Field Test" (25 min.)
DR. WILLIAM G. HOYT, System Development Corporation and
JAMES D. BAKER, U.S. Army Research Institute
Presented by: Dr. Hoyt

(Not Presented)

"The Application and Evaluation of Plato IV in Air Force Technical Training" (45 min.)
CAPT WILLIAM P. MOCKOVAK, AF
Lowry AFB;
CAPT PERRY S. MAIN, AF
Chanute AFB; and
GEORGE P. SCHARF, AF
Chanute AFB

Arabian Room
1315-1500

Personnel Assessment
Moderator:
LT CHARLES F. GULDENSCHUH, USCG

Wednesday Afternoon

Arabian Room

"Assessment and Counseling for Leadership Development" (20 min.)
DR. J. A. CAVINESS, U.S. Army Research Institute, Arlington, VA

"Dimensional Analysis of Leadership Measures from a Training Program" (20 min.)
RONALD G. DOWNEY, U.S. Army Research Institute
Presented By: Dr. Arthur C.F. Gilbert

"Assessment of Overseas Performance" (20 min.)
TED YELLEN, NPRDC

THURSDAY MORNING, OCTOBER 24

Palomino Room
0830-0930

Management and Measurement Problems
Moderator:
LTJG JAMES D. JONES, USCG

"Response Bias in Worker Task Inventories" (25 min.)
MAJ W. E. SLACK, USMC, Office of Manpower Utilization

"The Conflicting Demands of Go-No-Go Performance Criteria and the Requirements for Rank Ordering Students" (25 min.)
DR. WORTH SCANLAND, Naval Education and Training, Pensacola, FL

Appaloosa Room
0930-1000

Coffee

Palomino Room
1000-1200

Symposium: The Care and Feeding of Test Constructors
Chairman:
LT COL GEORGE D. OETTING, Extension Course Institute, Gunter AFS, AL
Comments by: LT COL Oetting

Thursday Morning

Palomino Room

"The Perspective of a Testing Psychologist Who Conducts Orientation Programs for New Test Constructors"
WALTER BIRDSALL, NETPDC, Pensacola, FL

"The Perspective of a Military Supervisor of Test Constructors Who Are Temporarily Brought in From the Field to Write Tests"
1LT DAVID VAUGHAN, USAF, Occupational Measurement Center, Lackland AFB, TX

"The Perspective of a Civilian Test Psychologist Working with Test Constructors"
CARL SMITH, USAEEC, Ft. Benjamin Harrison, IN

"The Perspective of a Non-Commissioned Officer Who is Actually Constructing Tests"
ADCM J. R. HARRIS, Coast Guard Institute, Okla. City, OK

"Results of a Service-Wide Survey of Enlisted Test Constructors"
LT COL OETTING

1155-1200 Announcements

1200-1315 Lunch

THURSDAY AFTERNOON, OCTOBER 24

Quarterhorse Room
1315-1445 General Session
Moderator:
LTJG ROBERT B. ELLARD

"The Development of a Student Evaluation of Course and Instructor Report" (25 min.)
LT COL G. J. CARPENTER, Royal Military College of Canada, Kingston, ONT

"Problem Checklists, A Pragmatic Approach for Fostering Unit Effectiveness" (25 min.)
FRED W. TIMMERMAN, Office of Military Leadership, West Point, NY
Presented by: CAPT RANDALL RIGBY
U.S. Army, University of Oklahoma

Thursday Afternoon

Quarterhorse Room

"Classification of Army Officer Duty Positions: Implications for Training Requirements" (15 min.)
DR. ARTHUR C. F. GILBERT, U.S. Army Research Institute, Arlington, VA

"U.S. Merchant Marine Licensing Examinations: The Modernization of a Traditional Mission" (25 min.)
MR. RICHARD C. WILLING and
LT DAVID MCDEVITT, USCG
U.S. Coast Guard Institute
Presented by: LT McDevitt

Palomino Room
1315-1445

Armed Services Vocational Aptitude Battery
Moderator:
LTJG TIMOTHY J. FLANAGAN, USCG

(Not Presented)
"ASVAB--What Is It? (60 min.)
DR. HARRY D. WILFONG, Armed Forces Vocational Testing Group, Randolph AFB, TX

(Not Presented)
"Large Group Testing" (20 min.)
CAPT MELVIN B. GAMBRELL, Armed Forces Vocational Testing Group, Randolph AFB, TX

1515-1630 Business Meeting

Cumberland Room
1630-1930 Social Hour

North Ballroom
1930-2200 Banquet

FRIDAY MORNING, OCTOBER 25

Palomino Room
0830-1030

Moderator: LCDR J.C. QUINN
Symposium: Evaluating Job Proficiency
Chairman:
DR. MILTON H. MAIER, Army Research Institute, Arlington, VA

"Functionally Integrated Performance Testing"
WILLIAM C. OSBORN, HumRRO, Div. 2

"Test Fidelity--Congruence of Test Tasks with Real-World Performance Tests"
DR. ANGELLO MIRABELLA and
EUGENE JOHNSON, Army Research Institute

Friday Morning

Palomino Room	"*Development of Criterion Ratings to Measure Job Performance*" CLIFFORD P. HAHN, American Institute for Research, Washington DC
	"*Scaling Techniques to Reduce Rater Leniency on Navy Enlisted Performance Evaluation Forms*" DAVID ROBERTSON and MARJORIE H. ROYLE, USN, NPRDC
Appaloosa Room 1030-1100	Coffee
1100-1200	Concluding Comments

MTA KEYNOTE ADDRESS

TITLE: EXTENDING THE COMMON BOND

RADM R. W. DURFEY

CDR Roughgarden,
Ladies and Gentlemen:

It is an honor and a privilege to be the keynote speaker for the 16th Military Testing Association Conference.

I wish to extend my warmest greetings to our friends from Canada, England, and Australia. I am most pleased to welcome representatives from the Marine Corps, the Navy, the Air Force, and the Army. Also, the representatives from other government agencies, from universities, and from the professional business community.

I wish to echo to you all the welcoming remarks of CDR Roughgarden. The Coast Guard, as a whole, is honored and proud to be able to host you here on this occasion. Some, perhaps many of you, were not previously aware that the Coast Guard had an installation, over 100 strong, here in mid-America. We are, as you can see. We appreciate our fine relationship with the FAA and those Coast Guard officers and men who are stationed here consider it a good assignment.

I have been told by one of your members that these MTA conferences are enjoyable and that they are attended by very intelligent people who think on the most interesting problems. I know from the program that there are four days worth of stimulating and exciting discussions on topics such as "The Care and Feeding of Test Constructors."

The first thing I did on being asked to be your keynote speaker was to review what other keynote speakers had presented. Interestingly, other keynote speakers confess to having done the same thing. Such a review has considerable value because it points to the continuing nature of some of our problems. Many of the keynote speakers have been concerned with managing personnel resources; they have been concerned, as I am, with the day-to-day problems which fill the large part of an operator's effort. The manager is required to make many decisions which have far reaching impact. Now, anybody in this type of situation is going to want the best possible information, tools, and procedures to help him in the decision process. Accordingly, your past keynote speakers have urged you to develop improved personnel management tools and procedures to help better manage our manpower assets. Let me quote Lieutenant General Simpson in his 1971 keynote address; he stated, "I come to you as a layman, asking you the experts in your fields and in your disciplines, to give us the tools to work with. This is what we need." Last year, Lieutenant General John W. Roberts said, "I ask you to provide us with the tools we need to satisfy these requirements. When you bring us the tools, tell us how to use them, and how not to use them, and make sure we listen." Lieutenant General Roberts goes on to remind us that, "remember, we can't always wait for the research effort to be perfectly organized, the right researcher to be found, the research effort to be conducted in a simon-pure environment. Very often, we need answers now." These quotes illustrate the position of managers. As George Allen of the Redskins says, "the future is now." The researcher, on the other hand, must admittedly insulate himself from such decision pressures until he has developed an adequate technical solution.

The basic antagonism between these positions must be understood and respected by researchers and operators alike. Such understanding, hopefully, will fill the researcher with a sense of urgency and the operators with a sense of patience.

I understand that in one of the early MTA conferences, serious consideration was given to a proposal to discontinue future MTA meetings. It was felt that most technical testing problems had been solved. Fortunately, the proposal was defeated. More recently, with budgets being very tight, the MTA has found itself with some problems in attendance. This situation is unfortunate but I am sure your Steering Committee will find adequate solutions. In spite of these pressures, the conferences continue because you, as individuals, have put out extra effort to host and attend these meetings - also, because you care enough to prepare papers that contribute to understanding and knowledge in the solution of personnel problems. Perhaps the most important reason the MTA conferences continue is the emergence of new and increasingly difficult problems which need to be solved. Both individually and collectively, you have recognized new problems and presented them at previous meetings for consideration and resolution. Today, the technical problems facing the scientific community in general and the behavioral sciences in particular appear to be overwhelming. Most recently, Justice Burger and the Supreme Court have determined that it is illegal to use tests or personnel procedures that have not been demonstrated to be valid; and also illegal to misuse those that are valid. A valid test is one that has been proven to be related to performance on the job. I know that demonstrating such validity is most difficult, especially if you add the problems of cultural bias and sex differences. However, proper validation is critical to the survival of many of our selection tests, classification tests, and advancement procedures. Validation of these instruments is both a technical and a real world problem which will require your best efforts. Thanks to many of you, the MTA, and to the test construction guidelines of the American Psychological Association, the armed services have been working for some time to make testing equitable and valid for all groups. I understand that the search for on-the-job performance measures pre-dates the first MTA conferences, that many psychologists, dressed in their best armour, have gone forth to seek and do battle with on-the-job criteria ---- only to return with corrected validities of Point 2 or Point 3. However, by using Class 'A' school performance as an intermediate criterion, corrected validities of Point 4, Point 5, Point 6, and even Point 7 have been achieved. The selection and classification tests, therefore, were found to be valid in predicting school performance. Also, it is possible to identify and eliminate those applicants with a high probability of school failure. Both testing and training profited from the relationship, and all selection and classification testing became tied to training. In March 1954, the American Psychological Association first issued the technical recommendations for psychological tests and diagnostic techniques. These were endorsed by the American Educational Research Association and the National Council on Measurement in Education. So as early as 1954, testing and training were closely tied together. Advancement-in-rate exams and correspondence courses were also destined for a similar close relationship.

The theme of this symposium, "Testing and Training, The Common Bond" also indicates the close relationship. However, the marriage between testing and training in the military has been based on the assumption that training was job related. Unfortunately, most school performance measures have not been found to have high correlation with operational job performance measures--such as enlisted performance marks. If we look closely, we can see that most school performance measures are achievement tests of the paper and pencil variety. Most selection and classification tests are aptitude or ability tests of the paper and pencil type. We know that paper and pencil tests that measure similar cognitive functions tend to correlate. However, the correlations between paper and pencil tests and on-the-job performance measures have been essentially worthless. As stated in the APA test standards, correlation analysis is the most preferred way of demonstrating validity. Now, there is no question that any personnel procedure that is selective must be proven to be equitable and valid. The burden of proof is on the user's back. Therefore, selection classification, training, advancement-in-rate procedures and exams, and any other selective procedures must be validated against on-the-job performance measures. In reviewing the agenda of this conference, I have found that over 50% of the papers are highly related to this validity problem.

Of particular interest is the symposium on Job Task Analysis for it may provide an alternative validation procedure. The APA test standard recognizes that content validity can be utilized. A test is valid if it can be demonstrated that the test content duplicates behavior and skills required on the job. Accordingly, job task analysis can be used to evaluate and validate school curriculas, school performance measures, advancement-in-rate exams, and selection tools which involve knowledge factors similar to those required on the job.

When training has been validated in this form, then it is job related and it is reasonable to accept those selection and classification tests which correlate with job related training. Advancement-in-rate exams whose content is highly job related will have high content validity. However, the content validity approach may not be sufficient for our own use. Evaluation of advancement-in-rate procedures, assignment procedures, and on-the-job training all are dependent upon measures of job performance as well as job content. Those of us who have been in a position of leadership and who have been responsible for evaluating performance know that it is possible to differentiate levels of performance. I know which of my subordinates are doing well and which are not. Even in fast moving situations such as football the coach knows and adjusts for the relative strengths and weaknesses of his players. So, from the manager's point of view, job performance can be evaluated. Why do marks tend to pile up at the high end of the marking scale? Why don't these official measures of job success relate well to other measures of success such as school performance measures? Hopefully, the symposium on evaluating job proficiency will address some of these problems. Maybe, we will gain insights into the techniques of measuring those performance differences which we all know exist.

Only if our selection, classification, assignment, and promotion instruments and procedures are effective and equitable can we begin to achieve an optimum in personnel utilization. Accordintly, those of you who are responsible for developing and improving these instruments and procedures now have the opportunity to make real contributions to effective management. I wish to stress that equity of these systems is not just a legal problem. It is a moral problem. It is also a human-relations problem and as you all know, it is a technical problem of extending the common bond to on-the-job performance. This should be within the state of your art.

On this note, I turn the forum over to you, the experts, to proceed with the tasks at hand. I regret that I will not be able to be with you after today, but your program is in good hands, and I wish you continued success. Thank you.

SECTION I

THE DEVELOPMENT OF TRAINING REQUIREMENTS

DEVELOPMENT AND APPLICATION OF AN ALGORITHM FOR TRANSLATING JOB INFORMATION INTO TECHNICAL TRAINING REQUIREMENTS

by

Major Wayne S. Sellman
Directorate of Personnel Plans
HQ United States Air Force
Washington, D.C.

&

Squadron Leader John W. K. Fugill
HQ Support Command
Royal Australian Air Force
Melbourne, Victoria

INTRODUCTION

One advantage of occupational analysis almost always mentioned in the educational and psychological literature is its use in designing training systems. Yet despite the widespread acknowledgment of the efficacy of selecting instructional objectives from up-to-date job information, only a few researchers (Smith, 1964; Smith, 1966; and Ammerman, 1966) have applied themselves to the actual task of developing procedures for translating job data into forms specifically designed to help individuals responsible for curriculum definition. Ammerman and Melching (1966) have pointed out that in the past, curriculum specialists have merely responded to instructional demands as they were brought to their notice. Instructional feasibility and the authority status of the information source thus became the dominant considerations.

With the development of high speed electronic computers capable of processing masses of occupational data, the application of applied multiple linear regression models to solve the course design problem is now well within reach. In the U.S. Air Force millions of dollars have been saved by eliminating training on tasks which occupational survey data have revealed to be performed by few workers (Christal, 1973). However, data concerning the probability of school graduates doing certain work activities in their first or second assignment is insufficient to make sound decisions concerning school curricula (Christal, 1970). Ammerman and Melching (1966) have observed that, in regard to the derivation of instructional objectives, the major premise is that "relevance" and "criticalness" are the factors of greatest significance. Hunter, Lyons, MacCaslin, Smith, and Wagner (1969) in discussing this problem concluded that although all the military services had clear instructions for the development of well-worded training objectives, little guidance has been given to insure that the objectives were relevant to the job performances toward which training was directed.

Air Force Manual 50-2, Instructional System Development (1970) provides instructions for the preparation of training standards within the U.S. Air Force. Using job data collected from occupational analyses (Morsh, 1964, 1969), training developers match the present course plan with the results of the job survey to determine how well the course aligns with actual work performance. For training task selection, two criteria are recommended: "Number of job incumbents performing a task" and "criticality of the task." Criticality refers to the relative importance of a task to mission success . . . . "inability to perform a task in a precise manner when the need arises could have a serious impact on the organization's mission capability." The U.S. Army Procedures Manual, Development of Instructional Systems (1964) lists ten selection criteria, several of which appear to be equally relevant for the determination of proficiency levels; these criteria are universality, difficulty, cruciality, frequency, practicability, achievability, quality, deficiency, retainability, and follow-on-training.

Christal (1970), using insights developed from a detailed examination of USAF occupational analysis data has defined a set of factors with a work performance orientation, i.e., probability of performance, task difficulty, skill perishability, frequency of inadequate performance, consequences of inadequate performance, probability of emergency performance, and training school effectiveness. On the first factor, probability of performance, information is gathered from representative samples of job incumbents through the administration of occupational surveys (Morsh, 1964, 1969). On the second factor, relative job difficulty, information has been gathered across over 100 USAF job specialties by numerous investigators (Lecznar, 1971; Mead, 1970a, 1970b; Mead and Christal, 1970; Fugill, 1972, 1973; McFarland, 1974) and from 35 Royal Australian Air Force job musterings (Fugill and Sellman, 1974).

In each study, given the definition of task difficulty as "time to learn to perform a task satisfactorily," and holding constant such factors as aptitude, training, and work experience, work supervisors rated lists of tasks using a 7-point relative scale (1 = least time; 7 = most time). The inter-rater reliability coefficients, derived using the Lindquist Formula (1953), for the various groups of judges were uniformly high-from .92 to .99. At this time, no definitive reports have been published on information collected on the other Christal (1970) factors.

Given definitive information on any set of criteria, the method for combining such data into curriculum decisions is not simple. Although some tasks are clearly good candidates for formal training, it is more difficult to come to decisions about borderline cases. For example, ordinarily a task would not be included in a course if the probability of a worker having to perform it is very low. On the other hand, if the task might have to be accomplished in an emergency and if the consequences of inadequate performance are serious and if the probability of being able to perform the task adequately without formal training is low -- then it certainly should be included in a formal technical training course.

Christal (1973) has indicated that the U.S. Air Force is actively striving to develop procedures for evaluating each task from an occupational survey on the factors described above, and to develop equations which can be applied to each task to determine how much emphasis should be given to it in the training course. This goal is admirable for large organizations such as the USAF, but in the absence of sophisticated computer support, small organizations will continue to find it difficult to translate occupational data into training statements. Perhaps one solution to this problem is the development and application of a simple yet rational algorithm which takes advantage of the systematic, scientific collection of job information; but one which doesn't demand complex computer processing. The purpose of this paper, therefore, is to discuss the rationale and function of one such algorithm developed by the Royal Australian Air Force (RAAF) in establishing the appropriateness of training curricula. The relevant job information used in the algorithm is obtained through the RAAF occupational analysis program (Sellman, 1973; Sellman and Fugill, 1973).

METHODOLOGICAL ISSUES

Training course design like any activity within the realm of the behavioral sciences is dependent on sound and reliable input if the output is to have merit. Thus, before addressing the algorithm itself, it is first necessary to mention some of the methodological issues prevalent in its development. In particular, attention will be focused on task statement specificity and comprehensibility, the selection of competent job informational judges, and the definition of task selection factors.

Specificity and Comprehensibility. In the pursuit of reliable information, the specificity of task statements appears to be a critical consideration. Ideally, a task statement should conform to the following elements as closely as possible (Fugill, 1972):

a. A single, specific action word,

b. nomenclature of equipment, materials, tools, and job performance aids,

c. an indication of the size and scope of the work activity,

d. an indication of degrees of precision required or tolerances permitted, and

e. a mention of any restrictive factors which contribute consistently to task difficulty.

The structural similarity of task statements so prepared to Mager's (1962) instructional objectives is obvious, and their ultimate translation into terminal training objectives by course designers is a relatively simple matter.

For comprehensibility, the common sense criteria are:

a. Use simple words as opposed to unnecessarily difficult ones,

b. use technical terms correctly for everyday job meanings, and

c. be clear and concise in expression.

Selection of Competent Judges. Consistent with a work - oriented approach, data are not sought from individuals who may be remote from, and insensitive to, the realities of the job situation. Accordingly, information is solicited from job incumbents and their immediate supervisors rather than second line superiors or training school instructors. As discussed above, research in both the USAF and RAAF suggests that these workers and supervisors can and do provide highly reliable information (Mead, 1970a, 1970b; Mead and Christal, 1970; Fugill, 1972, 1973; McFarland, 1974; and Fugill and Sellman, 1974). Accordingly, the rating inputs into the algorithm can be used with a great deal of confidence.

Definition of Task Selection Factors. Using Christal's (1970) ideas as a basic guide, data were captured on five of the eight factors he suggests as possible criteria for selecting tasks for inclusion in training courses. The factors used were operationaly defined as follows:

a. Task Difficulty - How much time is needed to learn to perform each task satisfactorily?

b. Skill Perishability - How much practice is needed for skills of this task to be satisfactorily remembered and retained?

c. Frequency of Inadequate Performance - How frequently would the task be performed unsatisfactorily regardless of type of training?

d. Consequences of Inadequate Performance - How serious are the consequences of inadequate performance in terms of destroyed materials, wasted time, and loss of lives/injuries?

e. Probability of Emergency Performance - How often must the task be performed under emergency conditions that require instant, effective action with no time for preparation and practice?

The inter-rater reliabilities for these five factors consistently have ranged from .70 to over .90 across a wide sampling of RAAF job musterings. Again, confidence can be placed in data gathered on these task selection criteria.

SIMPLE ALGORITHM DESIGN

To this point specific methods for determining what aspects of job performance need training have not been detailed. The ideal may be stated simply: "Select those aspects which the trainees have not already mastered;" but it is difficult to put into practice. The first RAAF algorithm was developed using occupational analysis data and selection criteria information supplied by senior work supervisors from the Clerk Supply mustering. While ratings from supervisors were captured on the five factors previously mentioned for tasks performed by most clerk supply airmen, the intercorrelations of the factors ruled use of some as inappropriate. Basically, the intercorrelations tended to confirm certain common sense hypotheses about the nature of work. For example, tasks judged to be extremely difficult were also judged to be highly perishable with a high frequency of inadequate performance. On the other hand, the probability of having to perform a task in an emergency was not a function of difficulty. Already, part of the algorithm was becoming clear. The high correlation existent between task difficulty, skill perishability, and frequency of inadequate performance suggested it was not essential to use all three factors. This of course left consequences of inadequate performance and probability of emergency performance for consideration but since clerk supply airmen didn't seem to perform any work in an emergency environment that factor also dropped from contention.

Thus, two factors (task difficulty and consequences of inadequate performance) remained for use in conjunction with the probability of task performance (percent job incumbents performing). Of course, a three - factor algorithm also made selection of tasks to be trained a much simpler affair than if all criteria had come into play. (This was also a very pleasing development for the investigators, we might add). Table 1 presents some sample clerk supply information considered in translating job information into training objectives.

TABLE 1

FACTOR INFORMATION - CLERK SUPPLY MUSTERING - USED TO DERIVE TRAINING STATEMENTS FROM THE RESULTS OF AN OCCUPATIONAL SURVEY

Tasks Performed by Airmen with 1-12 Months Experience	Percent Members Performing	Task Difficulty	Consequences of Inadequate Performance
1. Draft and type correspondence	52.3	3.9	3.9
2. Check catalogs to identify items of equipment when stock numbers unknown	43.2	3.5	4.3

3. Dispatch correspondence	40.9	2.8	4.0
4. File stores vouchers and documents	29.6	2.9	3.5
5. Post stores vouchers to accounts	27.3	3.9	4.2
6. Ensure display of appropriate notices	21.6	2.5	2.8

* 600 Tasks Listed in the Clerk Supply Job Inventory

The algorithm took the following form.

1. Tasks Performed by 10% or More Members	Minus	2. Non-Trainable Tasks	Equals	3. Relevant Trainable Tasks

NOTES

(1) With ten percent members performing set as the designated cut-off, the 120 qualifying tasks accounted for 75% of all work time. The remaining 25% work time comprised 480 infrequently performed tasks.

(2) Clerk Supply training specialists eliminated 35 tasks which were not "trainable" in the formal school environment, i.e., Item 6 in Table 1.

(3) The relevant "trainable" tasks numbered 85 and accounted for 62% work time.

Relevant Trainable Tasks	Minus	4. Task Difficulty < 3.0 or Consequences of Inadequate Performance < 3.0	Equals	5. Work Oriented Training Course Content

(4) Tasks were eliminated if the rated values for difficulty and consequences of inadequate performance were BOTH below 3.0. This ensured that tasks of high difficulty and high consequences were retained.

(5) The final list was comprised of 73 tasks. The orientation towards actual work performance ensures that the subsequent training will be relevant to the immediate needs of job supervisors.

CONCLUDING REMARKS

The most obvious reason for using such a job - performance algorithm is that training programs so designed provide the greatest assurance possible that the graduate will know how to do his job because it promotes maximum transfer of learning from the training situation to the job. A task orientation also enables the student to practice many times the procedures with which he will be expected to be familiar in the field.

In conclusion, two points made by Christal (1973) concerning the application of occupational data to training should be noted and considered. First, job analysts cannot simply provide job information or even training task lists and expect trainers to use them. They must explain how the data/lists can be used and also convince trainers that the data are "good." In some instances, trainers are reluctant to eliminate training or tasks even in the face of overpowering evidence that such a move would be cost - effective.

Second, a course curriculum cannot be designed solely from occupational survey data. Very few training specialists support the notion that training should be totally task oriented. The most common reason given for this is that task orientation does not provide the trainee with the fundemental knowledge needed to generalize his skills beyond those for which he was specifically trained. To the work - oriented list of tasks, training experts must add those knowledges and skills which are required to learn to perform the tasks satisfactorily. Once the tasks, skills, and knowledges to be trained have been identified, training specialists must go into the field to observe the number and sequence of steps required in the performance of each task to be included in the course.

SUMMARY

All educational institutions are plagued with problems of maintaining the relevance of what they teach. Because of the obvious, direct impact of the relevance of training on the efficiency of their own operations, the Military Services have pioneered the way toward finding methods to make sure that their training schools teach what will be needed. Even here, however, the traditional responsibility of education to "transmit the culture" has hindered the development of what is really needed - effective procedures for regular and systematic review of the needs for trained manpower and for translation of these needs into efficient training programs.

This paper has described the rationale of one quantitative procedure that hopefully will lead to the development of more systematic methods for designing efficient training courses and for maintaining their efficiency. Obviously, the quantification of data does not eliminate judgment from the training curricula design process, but it does make the application of judgment more rational.

REFERENCES

Air Force Manual 50-2. Instructional System Development. Department of the Air Force. Washington, D.C.: December 1970.

Army Procedures Manual. Development of Instructional Systems. Department of the Army. Ft Devens, MA: 1964.

Ammerman, H. L. Development of procedures for deriving training objectives for junior officer jobs. HUMRRO TR-66-3. Alexandria, VA: Human Resources Research Office, May 1966.

Ammerman, H. L. and Melching, W. H. The derivation, analysis, and classification of instructional objectives. HUMRRO TR-66-4. Alexandria, VA: Human Resources Research Office, May 1966.

Christal, R. E. Implications of Air Force occupational research for curriculum design. Paper presented to the Minnesota Research Coordinating Unit for Vocational Education Seminar: Process and Techniques of Vocational Curriculum Development. University of Minnesota, November 1970.

Christal, R. E. The United States Air Force Occupational Research Project. Paper presented in Symposium, The State-of-the-Art in Occupational Research and Development, Navy Personnel Research and Development Center, San Diego, CA., July 1973.

Fugill, J. W. K. Task difficulty and task aptitude bench mark scales for the mechanical and electronics career fields. AFHRL-TR-72-40. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, April 1972.

Fugill, J. W. K. Task difficulty and task aptitude bench mark scales in administrative and general career fields. AFHRL-TR-73-13. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, October 1973.

Fugill, J. W. K. and Sellman, W. S. Design of work - oriented task selection algorithms for the development of vocational training courses. Paper presented to the 9th Annual Conference of the Australian Psychological Society, Perth, Western Australia, Australia, August 1974.

Hunter, H. G., Lyons, J. D., MacCaslin, E. F., Smith, R. G., and Wagner, H. The process of developing and improving course content for military technical training. HUMRRO TR-69-9. Alexandria, VA: Human Resources Research Office, May 1969.

Lecznar, W. B. Three methods for estimating difficulty of job tasks. AFHRL-TR-71-30. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, July 1971.

Lindquist, E. F. Design and analysis of experiments in psychology and education. Boston, MA: Houghton - Mifflin Company, 1953.

Mager, R. F. Preparing instructional objectives. Belmont, CA: Fearon Publishers, 1962.

Mead, D. F. Development of an equation for evaluating job difficulty. AFHRL-TR-70-42. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, November 1970(a).

Mead, D. F. Continuation study on development of a method for evaluating job difficulty. AFHRL-TR-70-43. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, November 1970(b).

Mead, D. F. and Christal, R. E. Development of a constant standard weight equation for evaluating job difficulty. AFHRL-TR-70-44. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, November 1970.

McFarland, B. P. A comparison of task difficulty ratings made by nurses and medical service corpsmen. AFHRL-TR-74-7. Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory, March 1974.

Morsh, J. E. Job analysis in the United States Air Force. Personnel Psychology, 1964, 37, 7-17.

Morsh, J. E. Collecting, analyzing, and reporting information describing jobs in the United States Air Force. Paper presented to the 77th Annual Convention of the American Psychological Association, Washington, D.C., September 1969.

Sellman, W. S. Job analysis in the Royal Australian Air Force. Paper presented to the 8th Annual Conference of the Australian Psychological Society, Sydney, New South Wales, Australia, August 1973.

Sellman, W. S. and Fugill, J. W. K. Occupational analysis in the Royal Australian Air Force. Paper presented to the 15th Annual Conference of the Military Testing Association, San Antonio, Texas, October 1973.

Smith, R. G. The development of training objectives. HUMRRO Bulletin # 11. Alexandria, VA: Human Resources Research Office, May 1964.

Smith, R. G. The design of instructional systems. HUMRRO TR-66-18. Alexandria, VA: Human Resources Research Office, November 1966.

ONE OF THE MISSING LINKS

BY

COMMANDER B. CORMACK

CANADIAN FORCES

My purpose today is not to present an elaborate set of statistics or demonstrate a new and exciting breakthrough in training technology. First lets establish my credentials or lack of same; I am not an educational or training researcher. I am involved in occupational task analysis in the Canadian Forces, and what the occupational task analysis function is interested in is the employment of people in the Forces and the relevancy of their training to the work situation.

Now, when a statement such as this is made, trainers sometimes get very upset. Some feel we are trespassing on their preserves; we are "the new kids on the block" and how can we know anything about training? I don't think this attitude is prevalent now. An early contributing factor may have been the, shall we say, zealous method by which occupational task analysis was sold. To a degree we were guilty of the "sun syndrome", everything revolved around our function and those who thought they knew everything were particularly aggravating to those of us who did.

Those of us in task analysis are indebted to the pioneering work done by AFHRL, and particularly Dr. Christal and his associates, in developing the Task Inventory technique for application and computer analysis. A number of Services, outside your own are using the Comprehensive Occupational Data Analysis Program (CODAP) package to look at occupations, classifications, trades, MOSs, functions or whatever your particular Service calls the sub divisions of work effort that involves people. Where we differ however, is to me very significant and may be part of our difficulty in communicating analysis results to those who must react to the data.

Most task inventories are constructed to elicit two responses; performance and time spent performing. The time spent is an essential factor for analysis by CODAP. With few exceptions performance is confined to, and defined as, "Doing" the task.

As I said at the beginning of this presentation, I am not a training or educational researcher but the world of work that I see is not one of "Doing" or "Not Doing". It is peopled by those who "Do" low skill or high skill tasks, those who "Assist" (help) in low or high skill tasks, some who get involved "Doing and Supervising" tasks and some who "Supervise". Someone may have other definitions of skill involvement but, these appear to make sense to us.

Let me use an example; this slide represents responses to task statements where "Doing" only is a factor. People had the opportunity to respond in one of five ways but we extracted only those who are "Doing".

The next slide shows the responses graphed to display the mean involvement by rank with the respondents other four choices portrayed. Other than this slide being more visually pleasing, so what?

Let me digress for a moment. Picture, if you will, a serviceman confronted with responding to a task inventory. He reads or is told the instructions for completing it. He starts on a "Do/Not Do" basis. He reads a task; no, he doesn't actually do it but he knows he has something to do with it, does he respond? Maybe, maybe not. If, after going through the inventory he feels threatened (You mean, I only "Do" twelve tasks out of this lot?) he may rationalize that he is involved to some degree in other tasks so he'd better respond to those. How do I know this happens as described? I don't, but I have seen references in descriptions of occupational analysis programs where the need to call respondents is cited "due to a misunderstanding concerning the task" or, "more low skill level personnel responded to this supervisory task than did supervisors." Unfortunately, I haven't seen anything in the literature that explored these occurrences in more detail; the effect is known but not the cause.

I'll go out on a limb and say that I believe part of the cause is restricting the level of involvement response to a task.

Let me give another example to support this hypothesis. We administered a Task Inventory to the Vehicle Repair field. This inventory was administered to all maintenance levels; field units to overhaul and repair depots. There was a core of Corporals who stood out from the other Corporals because they had responded to the "Supervise" choice of many "hands on" tasks and "Doing" many "Management" tasks. Everyone knows that Corporals don't do those things so we looked a little deeper. We found they were all with our land forces in Germany. We also found they were at the intermediate maintenance facility and they actually supervised indigenous German labor (mechanics). The reason for the "unexpected" or "suspect" response was that the unit establishment was incorrect due to changes in maintenance policy. One wonders, what one response level would have produced.

I was interested in the methodology expressed for proficiency testing at last year's conference.[1] I believe the modular approach emphasizing the skill level is probably one of the better methods for determining proficiency. My only question in duty module construction is in it's validation. If the task analysis or data gathering phase obtains data on a "Do/Not Do" response, then you may lose something by management then saying this skill level "Supervises", "Does and Supervises", "Does" or "Assists" this task in this duty module. Why not start the analysis step with data reflecting levels of involvement? It would tend to reduce the subjective interpretation by management.

To return to the presentation of data, we have found that it is easier to demonstrate to training activities the skill relationships using the graphed format rather than the Job Description. It is more work to produce but it reduces the time involved in explanations.

One of the expressed or implicit uses of occupational data frequently stated is to establish the linkage between OJT and formal training programs. I am not aware of any research done in this area but again I'll go back to the graphed information.

By examining the degree of skill involvement a pattern seems to emerge. A more or less orderly progression appears. It is this examination that must be performed by subject matter experts; not educational specialists. The instructors who are knowledgeable concerning the occupational field can best determine the enabling objectives and terminal behaviours required to perform at a specified skill level.

I said I was going out on a limb before and I'll now saw it off. I believe that "Assisting" in the performance of tasks should form the heart of OJT packages. "Doing" tasks either requires exposure at the "Assist" level and/or formal training.

Careful examination of the tasks responded to at the "Assist" level may reveal that lower mental category personnel can be employed to advantage; theirs and the Service's. Repetitive involvement at this level may also enable them to progress on a building block basis to the more complex tasks. If this holds up then perhaps selection criteria for recruiting should be modified to tap a reservoir of neglected manpower.

---

[1] Stephenson, R.W., Davis, W.P., Hadley, H.I., and Cory, B.H., "A Modular Approach to Proficiency Testing" Proceedings 15th Annual MTA, October, 1973 Page 284 - 285

Do you want to test these hypotheses? I re-iterate that we in the Canadian Forces task analysis business are not research oriented, but it strikes me that you have the best vehicle to look at these areas if you want to explore them. The vehicle is your Inter Service Occupational Task Analysis Program, ISOTAP.

My understanding of ISOTAP is that a "common" task inventory is administered by each Service to it's people. The USN uses levels of involvement in it's data collection instrument. Perhaps you can prevail on the USN to give your particular Service some of it's response packets in exchange for some of yours when you administer the inventory. This would be a controlled study and I think you may be surprised by the responses you gather and subsequently analyse, if you try it.

STRUCTURAL TASK ANALYSIS--THE BRIDGE BETWEEN SELECTION AND TRAINING

James W. Dees

PRESENTED TO

MILITARY TESTING ASSOCIATION CONFERENCE
OKLAHOMA CITY, OKLAHOMA

21-25 OCTOBER 1974

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION
POST OFFICE BOX 428
FORT RUCKER, ALABAMA 36360

(205) 255-2712
Autovon 558-2712

ABSTRACT

Structural Task Analysis--The Bridge Between Selection and Training

James W. Dees

A procedure for using Guilford's structure of the intellect as the theoretical basis for a task analysis model is presented. It is reasoned that such a model would furnish a bridge between task analysis and test selection, and also a bridge between test selection and test validation. Such a mechanism might answer some of the EEOC criticisms of psychological testing because of the inherent content validity of the technique. In addition, the technique could be used to produce task analyses of both a job and training for that job. A comparison of the structural task analyses of the job and job training would expose discrepancies where abilities not required on the job are required for training. It might be then possible to modify training to bring the abilities required for training more in line with those required for the job, thus ultimately providing a bridge between task analysis through selection to training.

STRUCTURAL TASK ANALYSIS--THE BRIDGE BETWEEN SELECTION AND TRAINING

James W. Dees

Task analysis, task taxonomy, personnel selection, and the study of intelligence factors are all closely related. Good job descriptions are based on good task analysis and are helpful in choosing tests for personnel selection. The study of intellectual factors has very strong implications for the design of psychological tests for personnel selection. I believe that these four areas can be unified by a single system which will allow immediate translation from a task analysis based upon task taxonomy to the choice of personnel selection tests and to the intellectual requirements of the job. It would then be possible to redesign jobs and the training for those jobs in order to make the intellectual requirements consistent with the intellectual capabilities of the population of humans available to perform the jobs. I propose that this can be done by modifying a component theory of intelligence and reversing the order in which it is used in order to produce a task analysis technique in which the task is described primarily in terms of the intellectual components which it requires rather than in terms of the actions which are taken.

Paul Fitts (1961) presaged this when he said that "...a taxonomy should identify important correlates of learning rate, performance level, and individual differences." He emphasized the need for a taxonomy of processes and activities rather than of static elements. Describing the task in terms of its intellectual components is certainly a description in terms of processes and activities. Melton (1967) emphasized the need for a unifying theory in these areas in his statement that it was not enough to know only certain empirical relationships. Isolated empirical relationships are not normally generalizeable unless they are related to a unifying theory (Melton, 1967). Similarly, Miller emphasized the system's requirement of a taxonomy. He stated that the terms used in a taxonomy must be formally related to each other by explicit principles of inclusion and exclusion (Miller, 1962). A task analysis system based upon structured intellectual components would satisfy Miller's requirements. Fleishman and Stephenson (1970) delineated three different task taxonomic approaches:

1. the ability requirement approach,

2. the task characteristic approach, and

3. the system's language approach.

A task taxonomic approach based upon structural components of intelligence would look simultaneously at all three methods.

Several task taxonomic systems have already been developed. For example, Stolurow (1964) investigated several dimensions of task classifications. These dimensions have stimulated considerable research of value, but have not led to the design of a comprehensive task classification scheme. Gagne and Paradise (1961) developed an empirically derived hierarchy of learning sets for a specific type of mathematical task. While the learning sets themselves do not compose a task taxonomic system, they hint at the possible existence of such a system based upon the intellectual requirements which the job imposes. In fact, the learning sets identified were very similar to the elements in both Guilford's structure of the intellect and Cattell's intelligence and personality factors. Gagne further refined his thinking, and by 1965 had classified all types of learning into eight general categories: signal, stimulus response, chaining, verbal associates, multiple discrimination, concept, principle, and problem solving (Gagne, 1965). This is a very useful set of descriptors for educators, and has proven itself very valuable in the design of curricula. However, its scope is inadequate for use as a task taxonomic system. Bloom's taxonomy of educational objectives is similarly useful in the design of educational curricula (Bloom, 1956). However, a task analysis based upon this taxonomy would not result in an abstraction of task which would allow a mathematically meaningful comparison across different jobs.

In order to be maximally useful, the task taxonomic system must not only be a classification scheme, but must also bridge the boundary between the requirements of the job and the description of the man. In short, the task taxonomic system should be expressed in the language of the human ability system. Therefore, the current efforts in the field of intelligence testing are directly relevant to task taxonomy and task analysis. At present, there is a significant debate in the literature as to the validity of intellectual component systems as opposed to single factor systems. There is also debate among multifactor proponents as to the number of factors involved. I personally am not sure as to how this conflict will be resolved. However, the primary virtue of a theory is not whether or not it is true, but whether or not it is useful and produces valid results. The structure of the intellect as defined by J. P. Guilford has been attacked from many quarters as being an overextrapolation and an improper use of factor analysis. For example, Eysenck (1973) says "Guilford's scheme has been widely accepted because of its neatness and because of the tremendous amount of empirical work that has gone into it. It is unfortunate that it is not really acceptable on psychometric grounds." Eysenck's criticism of Guilford's work may ultimately prove correct. However, the complexity and diversity of Guilford's system offers the best available base for beginning the development of a structural task analysis system. It may be that much of Guilford's system is modified in the process. However, in the development of a new applied system, it is probably easier to eliminate and consolidate erroneous and duplicating categories than it is to hypothesize and validate new categories. For this reason, I propose to use Guilford's structure of the intellect as the starting basis for the development of a structural task analysis system.

AN OUTLINE OF THE STRUCTURAL TASK ANALYSIS SYSTEM

Slide 1 in Appendix 2 of the handout presents the familiar Guilford model of the structure of the intellect. Here you see the three dimensions: operations, products, and content. Each of these divisions is further divided into subdivisions. The combination of various subdivisions from each of the three dimensions produces a total of 120 separate elements within the design. Not all of these elements have been identified empirically (Guilford, 1968). In addition, the tests developed in attempts to measure individual elements of the structure have failed to achieve a zero correlation with one another (Dees, 1972). Nonetheless, where the structure may have failed in theory, it has succeeded in practice. In a study by Dees (1972), 36 separate Guilford elements were tested and correlated against three different criteria. The three criteria were:

1. ability to shoot the M-16 rifle,
2. the peer rankings of students in Infantry Officer Candidate School (OCS), and
3. whether or not an OCS student completed the course.

In each case, the contribution of each of the individual tests was quite small, but the intercorrelations between the tests were also low and the resulting multiple correlations with the three criteria were relatively high: .44, .41, and .86, respectively, after correction for shrinkage. This is the sort of situation which would be expected using a large number of fairly independent measures, each of which makes a small but relatively independent contribution to the prediction of a complex skill. In short, although Guilford's theory may not be all that Guilford claims it to be, it can be used to insure heterogeneity of test material in a multivariate format. This approach to structural task analysis assumes only that Guilford's work has provided a panoramic presentation of intellectual skills and tests for those skills. While I am not assuming that Guilford's system is theoretically correct, I am saying that the Guilford system may ultimately be of more practical usefulness because of its heterogeneity and comprehensiveness than another theory which might more closely parallel the abstract qualities of man's intellect.

Guilford and Hoepfner (1963) have defined each of the subdivisions of his structure of the intellect. Many of these definitions require modification in order to be applicable to a task analysis situation. A first cut at these redefinitions is furnished in Appendix 1 to the handout of this paper. Structural task analysis then differs from traditional task analysis in the nature of its descriptors. In traditional task analysis, it is the job itself that is being described. In structural task analysis, the job will be described traditionally for identification purposes, but the heart of the system will be a description of the intellectual requirements for each job component based upon a modification of Guilford's structure of the intellect.

USES OF THE SYSTEM

There are two ways in which the results of a structural task analysis can be used. First, separate tests can be constructed for a wide variety of jobs. The tests would contain items representing the intellectual elements identified in the structural task analysis, and would contain them in proportion to their relative importance. Second, a single massive testing program could be developed which would identify an individual's entire spectrum of capabilities and would enable that individual to be provided with a listing of all of the jobs for which he has untrained capability. Furthermore, the listing could provide this information in the order of descending capabilities. The individual would then be selected according to his capabilities in all of the ability areas important for a job with the weighting of those areas identical to the weighting of the importance of the job itself. In short, this would be an improved version of the classification battery where the tests administered should have predictive capabilities for virtually every imaginable job.

In order to develop and prove this system, it must be used and validated traditionally on a complex and well-defined job. The job must be complex so that a wide variety of intellectual elements can be examined. It must be well defined in order to reduce error variation due to disagreement as to the nature of the job. It must be validated traditionally in order to demonstrate its consistency with conservative statistical methodology. However, once validated using traditional methodology, this technique would be applicable to jobs which heretofore have defied traditional criterion validation. It should be possible to identify the intellectual elements required by a given job whether or not the criteria of good job performance can be measured quantitatively. Thus, it should be possible to apply this technique to management positions where criteria of performance are qualitative, and to combat positions where the obtaining of criteria of performance can be dangerous. In most systems, the analysis of the task, the selection of the test instrument, and the validation of the test instrument are separate steps. Once structural task analysis has been validated, all three of these steps will then be woven into one process.

PERSONNEL POLICY CONSIDERATIONS

No treatment of intelligence testing or personnel selection is complete without a consideration of the interaction of those issues with the current personnel policy decisions of the Federal Government and its courts. The Equal Employment Opportunity Act of 1972 is the legal basis for court requirements that businesses across the nation validate their selection instruments. All military services are currently under the scrutiny of Congress to observe personnel selection and allocation programs at least as equitable as those observed by civilian organizations. The Equal Employment Opportunity Commission (EEOC) created by the 1964 Civil Rights Act is charged with the responsibility of administering the law. In order to do this, it has developed a series of informal publications known as guidelines. While these guidelines do not have the force of the law, they are an interpretation of the law which the courts and Congress have in the past tended to uphold. The publications most pertinent to employment are:

1. *Guidelines on Employment Testing Procedures* published in 1966,
2. *Guidelines on Employee Selection Procedures* published in 1970,
3. *Guidelines on Discrimination Because of Sex* published in 1972.

Of these, the most important is the *Guidelines on Employee Selection Procedures*.

Section 1607.2 of the selection procedures guidelines defines what is meant by a test. This definition includes any paper and pencil or performance measure used as a basis for any employment decision. Thus, according to the EEOC there is no employment selection device of any sort which is exempt from the law. Any personnel selection program developed by the proposed program would eventually be examined to determine its conformance with the requirements of the law.

In Section 1607.4 the employer is required to have available for inspection evidence that the tests do not discriminate in violation of Section 1607.3. Where differential rejection rates for minority and nonminority candidates are in evidence, the employer must: (1) provide evidence of the test's validity, and (2) provide differential validity for the minority groups where it is possible.

The requirement for differential validity has led to a great deal of discussion. One set of research indicates that there is no such thing as differential validity. This view holds that while there may be differences in the mean predicted performance of various racio/ethnic groups, these differences in predicted performance are borne out by differences in actual job performance. Another group of researchers have found contradictory evidence indicating that racio/ethnic group differences in actual job performance are not validly indicated by racio/ethnic group differences in test prediction of that job performance.

Differential validity is a sticky question for two reasons. First it is extremely expensive. Second, many people have an emotional negative reaction to any system that would allow members of one racio/ethnic group to be hired with a given selection test score, while requiring a higher selection test score for members of another racio/ethnic group.

The test selection technique proposed for this effort will not eliminate all of the cross cultural differences in test sophistication which in part account for the requirement for differential validity. However, if successful, this system will define objectively and systematically what the job requirements are. Further, these job requirements will be defined in a language which will permit their mathematical manipulation and comparison across very different jobs. This in itself is not a cure for the current alleged discriminatory practice in personnel selection testing. However, it is an avenue of approach to the problem.

IMPLICATIONS FOR TRAINING

Structural task analysis defines the job in terms of its intellectual requirements rather than the task per se. The descriptors used in structural task analysis also translate immediately into requirements for tests for that job. At the same time, the manner in which the analysis is conducted and the manner in which the tests are selected is intrinsically a form of content and construct validation. Thus, structural task analysis is simultaneously a bridge between task analysis and test selection, and a bridge between test selection and validity.

The selection of the best available people for a given job is itself a significant contribution to training. However, structural task analysis can ultimately be of much greater value in training. How often have you heard someone say that a given student would be an excellent pilot if he could just get through training, or that "This fellow may have graduated with honors from college but he doesn't know anything about the business." One of the implications of such statements is that the requirements for the completion of the training program are not consistent with the requirements for the job. This is not unusual, and in some cases it is unavoidable. The learning of material often places ability requirements on an individual which the application of that material in a job setting does not. However, the learning requirements placed upon the student are often a function not so much of the material which must be learned, but of the method of material presentation. Ideally, the training for a job should require abilities as close to the ability requirements of the job as possible. To the extent training deviates from this, students will be failed who could do the job well and will be passed who will fail on the job. Structural task analysis of the job can delineate the ability requirements of the job. Structural task analysis of the training program can delineate the ability requirements for the program. Discrepancies between the two analyses can be noted, and the training program can be changed in order to reduce these discrepancies.

SUMMARY

Structural task analysis, by furnishing bridges between task analysis, test selection, test validation, and training can furnish a unifying theory with which many improvements in all of these areas can be achieved. The improvements should reduce the inequity in job assignments among different ethnic groups. It should also provide a means for developing selection batteries for jobs whose criteria of performance are qualitative, dangerous, or otherwise difficult to measure. In addition, it should provide a means for improving the training for various jobs by decreasing the discrepancy between the ability requirements of training and the ability requirements of the job.

REFERENCES

Bloom, B. S. (Ed.) *Taxonomy of Educational Objectives--Handbook 1: Cognitive Domain*, David McKay Co., New York, 1956.

Dees, James W. *Complex Intellect vs the IQ Test as a Predictor of Performance*, Paper for Mid-South Educational Research Association Convention in New Orleans, Louisiana, 1972.

Eysenck, H. J. (Ed.) *The Measurement of Intelligence*, Williams and Wilkins Company, Baltimore, 1973.

Fitts, Paul M. Factors in Complex Skill Training in: Glaser, Robert (Ed.). *Training Research and Education*, Department of Psychology, University of Pittsburgh, 1961, 199-222.

Fleishman, Edwin & Stephenson, Robert W. *Development of a Taxonomy of Human Performance: A Review of the Third Year's Progress*, American Institutes for Research, TR-3, September 1970, 68 pp.

Gagne, Robert M. *The Condition of Learning*, Holt, Rinehart & Winston, New York, 1965.

Gagne, Robert M. & Paradise, Noel E. "Abilities and Learning Sets in Knowledge Acquisition," *Psychological Monographs: General and Applied*, 1961, *75*, (14) Whole No. 518.

Guilford, J. P. *Intelligence, Creativity and Their Educational Implications*, British Printing Corporation, Robert R. Knapp, Publisher, 1968, San Diego, California, 92107.

Guilford, J. P. & Hoepfner, R. *Current Summary of Structure of Intellect Factors and Suggested Tests: Studies of High-level Personnel*, Psychological Laboratory Report, The University of Southern California, December 1963.

Melton, Arthur W. Individual Differences and Theoretical Process Variables: General Comments on the Conference. In: Gagne, R. M. (Ed.). *Learning and Individual Differences*. Charles E. Merril Books, Inc., Columbus, Ohio, 1967, pp 238-252.

Miller, Robert B. Task Description and Analysis. In: Gagne, R. M. (Ed.). *Psychological Principles in System Development*, Holt, Rinehart, & Winston, New York, 1962.

Stolurow, Lawrence M. *A Taxonomy of Learning Task Characteristics*, Aerospace Medical Research Laboratory Report AMRL-TDR-64-2, Jan. 1964, 203 pp.

APPENDIX 1

REDEFINITIONS OF GUILFORD'S ELEMENTS OF THE INTELLECT FOR TASK ANALYSIS

Operations

Cognition: A cognition is a discovery, a rediscovery, or a recognition. In short, it is the perception of the existence of something. Quite often a cognition is the first intellectual step in a series required for problem solution. As such, it is a "perception" as compared to a "sensation." The important point is not that a stimulus impinges upon the individual, but rather that the individual perceives the stimulus even though it may be camouflaged by many other stimuli and recognizes its importance.

Memory: The memory classification is used for the commission of memory only. It is not used for the retrieval of information from memory even though the success or failure of the commission is not known until retrieval is required.

Divergent Production: Divergent production is tantamount to inductive logic. Whenever one generalizes from the specific case to the general rule, divergent production or inductive reasoning is taking place.

Convergent Production: Convergent production is tantamount to deductive logic. That is, whenever one is reasoning from the general rule to the specific application, convergent production or deductive reasoning is occurring.

Evaluation: Evaluation is the qualitative or quantitative judgment as to the suitability, goodness, correctness, or adequacy of what we know, remember, or produce intellectually.

Content

Figural: Figural content is concrete material such as is perceived through the senses. It does not represent anything except itself. Visual material has properties such as size, form, color, location, or texture. Things we hear or feel provide other examples of figural material.

Symbolic: Symbolic content is composed of letters, digits, and other conventional signs usually organized in general systems such as the alphabet or the number system.

Semantic: Semantic content is in the form of verbal meanings or ideas.

Products

Unit: A unit is an isolated element of information which has importance in and of itself. It may or may not be associated with a higher order system or class or relation. The important point is that in the intellectual process it is the unit that is created not a class of units or a relationship between units or a system of units.

Class: The production of a class occurs when two or more concepts are identified as belonging in the same category because of shared characteristics. For example, a baseball, a football, and basketball might all be classed as game equipment.

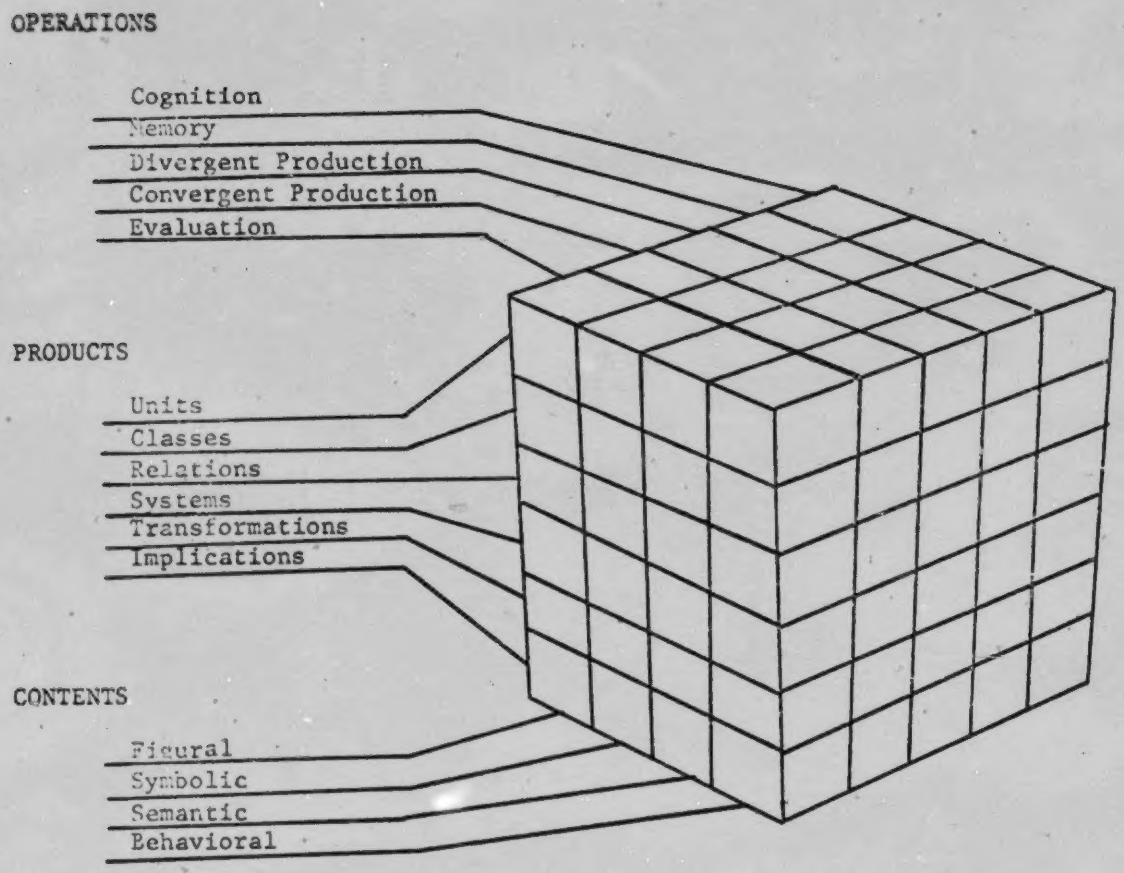
Relation: The production of a relation is similar to the production of a class in that it is based upon shared characteristics. However, the shared characteristics rather than being used to place two or more concepts in a single category are used to establish a relationship between two or more concepts. For example, it might be noted that multi-engine aircraft tend to be larger in volume and weight than single-engine aircraft. Thus, a relationship could be established between the number of engines an aircraft has and the probable size and weight of that aircraft.

System: A system is similar to a relation in that it is based upon the establishment of certain relationships. However, in a system the concepts are abstracted so that they apply to a greater population of situations. For example, instead of stopping at the relationship between the number of engines and the size of the aircraft, the relationship of these factors to the principles of lift and thrust might have been made. If this step had been taken, instead of a mere relationship between number of engines and size of aircraft, a system would have been established relating the size of the aircraft to its power requirements.

Transformation: A transformation is produced when the set or context of the information is changed. For example, a screwdriver may be used as a lever, or a double entendre is achieved with one wording.

Implication: An implication is an extrapolation of information to a probable end result or effect.

APPENDIX 2



OPERATIONS

Cognition
Memory
Divergent Production
Convergent Production
Evaluation

PRODUCTS

Units
Classes
Relations
Systems
Transformations
Implications

CONTENTS

Figural
Symbolic
Semantic
Behavioral

Slide 1

SECTION II

PERSONNEL ASSIGNMENT USING INTEREST INVENTORIES

Interpretive Report for the

Strong-Campbell Interest Inventory

Dr. Charles B. Johansson

National Computer Systems

Traditionally, reports formats for mental testing instruments present a series of scales and corresponding score values which are subsets of summated weighted responses of the total inventory. In some cases, the reporting format is a profile which graphically displays the magnitude of the individual scale score using a standardized scaling technique across all scales. Typically, interpretation of these scores necessitates an interview session with a professionally trained counselor or psychologists to explain the score meaning and give additional sources of information.

With increased complexity of tests and increased research on the instruments, the computer has been used more and more in the last few years to retrieve relevant information to aid in the interpretation of test results, such as interpretive results for Minnesota Multiphasic Personality Inventory (MMPI, 16 Personality Factors (16PF), and Strong Vocational Interest Blank (SVIB).

In July 1974, the Strong-Campbell Interest Inventory (SCII) became available commercially. This instrument is basically a combination of the male and female forms of the SVIB into a single inventory; now all individuals respond to a single inventory and everyone receives scores on all available scales. Another major innovation for the SCII is the structuring of scales around a theoretical occupational framework, specifically, that of John Holland's hexagonal model (Holland, 1973).

In August 1974, NCS' computerized interpretive report became available for the SCII. This report provided a service that would contain sufficient information for the basic and general understanding of the different types of scales and provided additional references and sources of information that could be used to further the person's understanding of the occupational worlds.

Some advantages in using an interpretive service are:
1) Counselors or other professionally trained psychologists may choose to release results to the client; narrative information similar to that presented by the counselor will appear on the interpretive output. The probability that the client will retain the important aspects of the meaning of the scores should be increased since easy reference can be made to the verbage in the report.

2) Relevant codes and page references for the Dictionary of Occupational Titles (DOT) and the Occupational Outlook Handbook (OOH) are presented for the person's high and low scores. In addition, relevant books and appropriate OOH reprints are referenced as further sources of information; thus, the counselor and client have to spend less time searching for these references. This could be expanded easily to include references such as armed services codes. Again, the probability that these materials will be used is increased since specific sources of information are presented.

3) The interpretive output provides a basic understanding of the results with cautionary statements against an over-interpretation. This would be important is situations where individual counseling may not be feasible, such as a large-scale research project where the results are returned to subjects as incentive for participating.

The interpretive computerized output for the SCII is not intended to replace the role of the counselor, but to provide explanatory comments for the test taker to aid in understanding scores on the SCII and the individual's interests. Basically, each computerized interpretive output reports five types of scores with statements explaining their meaning: 1) scores on six General Occupational Themes; 2) scores on 23 Basic Interest Scales; 3) scores on same-sex Occupational Scales; 4) scores on opposite-sex Occupational Scales; and 5) scores on Administrative Indices and Non-Occupational Scales.

INTERPRETIVE OUTPUT -- INTRODUCTORY COMMENTS

Reference to Appendix 1 which contains an example of a complete interpretive report for the SCII will be helpful in establishing the juxtaposition of many of the following sections.

The first page of the computerized interpretation is illustrated in Figure 1 below. Heading types of information, such as name, date of the report, and sex of the subject are listed.

Insert Figure 1 About Here

GENERAL OCCUPATIONAL THEMES

Development of SVIB General Occupational Theme scales during the last couple of years was extended to the SCII. These theme scales complement the structures of these inventories by providing a higher level of aggregation of the underlying basic dimensions of the inventories. A direct analogy with a hierarchical theory of personality is now possible where the Theme Scales are analogous to the very general Type Level, the Basic Interest Scales are analogous to the more specific Trait Level, and the Occupational Sclaes are analogous to the Specific Response Level.

FIGURE 1

EXAMPLE OF INTERPRETIVE SCII REPORT

001 0003

STRONG-CAMPBELL INTEREST INVENTORY

TEST RESULTS FOR NO NAME 9/12/74

THESE TEST RESULTS ARE FROM DATA YOU PROVIDED WHEN YOU ANSWERED THE STRONG-CAMPBELL INTEREST INVENTORY. THE RESULTS ARE BASED ON YOUR LIKE AND DISLIKE ANSWERS TO THE INVENTORY AND THEY ARE IMPORTANT IN HELPING YOU TO BETTER UNDERSTAND HOW YOUR PREFERENCES FIT INTO THE WORLD OF WORK. THEY POINT OUT WORK AREAS WHERE YOUR INTERESTS MATCH THOSE OF EMPLOYED PEOPLE. THEY ALSO CAN BE USEFUL IN INDICATING POTENTIAL PROBLEMS WHERE YOUR INTERESTS DIFFER SUBSTANTIALLY FROM PEOPLE IN OCCUPATIONS THAT YOU MIGHT BE CONSIDERING.

IMPORTANTLY, THESE RESULTS ARE MEASURES OF YOUR INTERESTS AND NOT OF YOUR ABILITIES. FOR EXAMPLE, THE SCORES MAY INDICATE THAT YOU WOULD LIKE THE DAILY ROUTINE OF ARTISTS AND THAT YOU LIKE ARTISTIC ACTIVITIES, BUT THEY WILL NOT TELL YOU IF YOU HAVE THE TALENT TO BE AN ARTIST.

THE RESULTS CAN GIVE YOU SOME USEFUL INFORMATION ABOUT YOURSELF, BUT DO NOT EXPECT MIRACLES. ALTHOUGH YOU MAY FEEL THAT THE SCORES ACTUALLY TELL YOU NOTHING MORE THAN YOU ALREADY KNEW, THEY WILL TELL YOU HOW HIGH OR LOW YOUR INTERESTS ARE IN RELATION TO THE AVERAGE SCORES OF OTHERS.

* *

THREE MAIN SETS OF SCORES ARE PRESENTED BELOW. THE FIRST SET OF SCORES ARE YOUR RESULTS ON SIX GENERAL OCCUPATIONAL THEMES--THEY GIVE YOU A GENERAL OVERALL VIEW OF YOUR INTERESTS. THE SECOND SET OF SCORES ARE YOUR RESULTS ON THE BASIC INTEREST SCALES--THESE SCORES TELL YOU ABOUT THE STRENGTH OF YOUR INTERESTS IN MORE SPECIFIC AREAS SUCH AS SALES, MUSIC, AND MECHANICAL. THE THIRD SET OF SCORES ARE THE OCCUPATIONAL SCALES--THEY TELL YOU HOW SIMILAR OR DISSIMILAR YOUR INTERESTS ARE TO EMPLOYED PEOPLE IN VARIOUS OCCUPATIONS SUCH AS ARTISTS OR MATHEMATICIANS.

* *

Theme Scales, patterned after Holland's Theory, were developed first for the SVIB during the early 1970s (Campbell and Holland, 1972; Hansen and Johansson, 1972), were revised for the SCII, and now are an integral part of the reporting schema. Holland (1966) reviewed the plethora of information that had been accumulated on vocational choice, stability, and achievement, and drew upon his clinical experience to develop his theory that the occupational world is represented primarily by six styles or types: REALISTIC, INVESTIGATIVE, ARTISTIC, SOCIAL, ENTERPRISING, and CONVENTIONAL. Each type involves a number of personal attributes and characteristics that predispose a person to enter a vocational field that is congruent with her or his individual style.

When Holland's theory was mapped into the SVIB, separate scales for males and females were developed for each of the six typologies. Essentially, 20 SVIB items were selected for each scale on the basis of a theoretical fit into Holland's model. For example, the items "Architect," "Astronomer," "Author of Technical Books," "Chemist," and "College Professor" all appear on the Investigative theme scale (see Table 1). "Like," "Indifferent," and "Dislike" responses to the items on the theme scales were assigned scoring weights of +1, 0, and -1, respectively. Thus, high scores on a theme scale are obtained by responding "Like," and conversely, "Dislike" responses result in low scores.

Insert Table 1 About Here

Each theme scale was normed on a general reference sample of 600 people (300 males, 300 females) who completed the SCII. This general sample has a mean of 50 and standard deviation of 10 on each of six scales; however, male and female distributions do differ considerably for some scales and their differences will be discussed later as they pertain to reporting of results.

Theme Scale Profile

On the interpretive report, a brief introduction of these scores is presented first on page 2 (see Figure 2). Statements are designed to emphasize that these scales are general measures of one's interests and that an average person in the general population scores between 43 and 57 on each scale. This range (43-57) represents the middle 50 percent of scores in the general reference sample.

Insert Figure 2 About Here

Scores on the theme scales are presented in profile form and later in

TABLE 1

EXAMPLE OF ITEMS ON INVESTIGATIVE SCALE

ITEM NUMBER	ITEM WEIGHTS L	I	D	ITEM
3	+1	0	-1	ARCHITECT
8	+1	0	-1	ASTRONOMER
13	+1	0	-1	AUTHOR OF TECHNICAL BOOKS
27	+1	0	-1	CHEMIST
33	+1	0	-1	COLLEGE PROFESSOR
67	+1	0	-1	INVENTOR
94	+1	0	-1	PHYSICIAN
103	+1	0	-1	PSYCHOLOGIST
112	+1	0	-1	SCIENTIFIC RESEARCH WORKER
121	+1	0	-1	SURGEON

FIGURE 2

EXAMPLE OF OCCUPATIONAL THEME PROFILE

GENERAL OCCUPATIONAL THEMES

PSYCHOLOGICAL RESEARCH HAS SHOWN THAT INTERESTS CAN BE GROUPED INTO SIX CATEGORIES AND THAT THESE GROUPINGS CAN BE DESCRIBED BY A GENERAL THEME. IMMEDIATELY BELOW ARE YOUR SCORES ON THE SIX GENERAL THEMES AND A GRAPH TO VIEW YOUR RESULTS BETTER. AN AVERAGE PERSON IN THE GENERAL POPULATION SCORES BETWEEN 43 AND 57 ON THE SCALES. SCORES BELOW 43 GENERALLY ARE CONSIDERED AS LOW AND SCORES ABOVE 57 GENERALLY ARE CONSIDERED AS HIGH.

```
          YOUR      3         4         5         6         7
SCALE     SCORE ....0.........0.........0.........0.........0....

R-THEME   77    *************************************************   (REALISTIC)
I-THEME   70    *********************************************       (INVESTIGATIVE)
A-THEME   50    *************************                           (ARTISTIC)
S-THEME   40    ***************                                     (SOCIAL)
E-THEME   30    *****                                               (ENTERPRISING)
C-THEME   20    *                                                   (CONVENTIONAL)
```

ALTHOUGH DESCRIPTIONS ARE GIVEN BELOW FOR EACH THEME, THEY DESCRIBE A GROUP OF PEOPLE AND ALL THE CHARACTERISTICS MAY NOT FIT EXACTLY ANY ONE PERSON. FEW PEOPLE SCORE HIGH ON JUST ONE THEME AND LOW ON ALL OTHERS. MOST SCORE HIGH ON TWO OR EVEN THREE THEMES AND THUS THEY SHARE SOME OF THE CHARACTERISTICS WITH MORE THAN ONE TYPE. SOME PEOPLE SCORE LOW ON ALL THEMES WHICH WOULD INDICATE THAT DEFINITE PATTERNS OF INTERESTS, AS MEASURED BY THESE THEMES, HAVE NOT BEEN ESTABLISHED YET--THIS IS PARTICULARLY TRUE FOR YOUNG PEOPLE. GENERALLY, THE HIGHER YOUR SCORE, THE MORE CHARACTERISTICS YOU SHARE WITH THAT PARTICULAR THEME. MEN AND WOMEN SCORE SOMEWHAT DIFFERENTLY ON SAME THEMES AND THIS HAS BEEN TAKEN INTO ACCOUNT BY COMPARING YOUR SCORE WITH THE AVERAGE FOR YOUR SEX.

an analysis section where psychological constructs of the scales are discussed. The scores on the theme scales are ordered from highest to lowest with a graphical representation of each value as indicated in Figure 2.

The left side of the profile uses letter abbreviations for the theme scales (R,I,A,S,E, and C), while the right side contains the expanded name. This abbreviation schemeis emphasized throughout the interpretive report, thus, an R-I-A-S-E-C concept is stressed.

Theme Scale Analysis

Descriptive phrases are given for each of the six General Occupational Theme scales and analyses of the score values are located on the lower one-half of page 2 through page 4 of the interpretive report. Order of presentation is identical to the profile representation -- highest score first and so forth. Descriptive phrases for the six types are based on the work of Holland. They represent the extreme model for each of the scales. Again, the scales are references primarily by their mnemonic abbreviations of R-I-A-S-E and C Theme. [Additional information the General Occupational Theme Scales can be vound in Campbell (1971, 1974), Campbell and Holland (1972), Hansen and Johansson (1972), Holland (1973), Johansson (1971), and Roberts and Johansson (1974)].

Following each descriptive paragraph on the interpretive report is a one or two line sentence that starts: "YOUR SCORE OF xx" where xx is the person's score on that particular scale; the remainder of the sentence is dependent upon the actual score value and gives the person an estimate of the relevancy of the descriptive phrases. Figure 3 exemplifies the possible statements and corresponding score values that are related to the R-THEME descriptive paragraph. Only one of these statements will appear for the R-THEME and that statement is dependent upon the person's scores. For example, a score of 50 on R-THEME falls in the range of 43-57 and the middle statement is printed out after the R-THEME description.

Insert Figure 3 About Here

The above score ranges and corresponding narrative statments relate the individual's score to that of the general reference sample. Another level of complexity is added after these phrases; each score is related to the distribution for the individual's sex.

An interpretive comment is presented that indicates the relative strength of the person's interest in comparison to the individual's sex. Phrases of "VERY HIGH to "VERY LOW" that were dependent upon score distribution for the respondent's sex were used differentially for each of the scales. Each statement commences with either "IN COMPARISON WITH OTHER

FIGURE [illegible]

INTERPRETIVE STATEMENTS FOR REALISTIC SCALE

SCORE	STATEMENTS
-34	YOUR SCORE OF xx ON THE R-THEME SCALE INDICATES THAT MOST LIKELY NONE OF THE ABOVE DESCRIPTIONS ARE TRUE OF YOU AND YOU WOULD NOT ENJOY ACTIVITIES WHERE YOU WORK WITH YOUR HANDS.
35-42	YOUR SCORE OF xx ON THE R-THEME SCALE INDICATES THAT MOST LIKELY VERY FEW OF THE ABOVE DESCRIPTIONS ARE TRUE FOR YOU AND YOU MIGHT NOT ENJOY ACTIVITIES WHERE YOU WORK WITH YOUR HANDS.
43-57	YOUR SCORE OF xx ON THE R-THEME SCALE INDICATES SOME OF THE ABOVE PHRASES ARE PROBABLY TRUE FOR YOU, WHILE SOME OF THE OTHER PHRASES PROBABLY DO NOT DESCRIBE YOU AS WELL.
58-65	YOUR SCORE OF xx ON THE R-THEME SCALE INDICATES THAT VERY LIKELY MANY OF THE ABOVE DESCRIPTIONS ARE TRUE FOR YOU AND YOU MIGHT ENJOY ACTIVITIES THAT ALLOW YOU TO WORK WITH YOUR HANDS.
66+	YOUR SCORE OF xx ON THE R-THEME SCALE INDICATES THAT VERY LIKELY MOST OF THE ABOVE DESCRIPTIONS ARE TRUE FOR YOU AND YOU WOULD ENJOY ACTIVITIES THAT ALLOW YOU TO WORK WITH YOUR HANDS.

FEMALES..." or "IN COMPARISON WITH OTHER MALES..." Completion of the sentence depends on the score and sex of the individual. Table 2 lists the score range by sex and the corresponding descriptive phrase. For example, a score of 49 for a male on the R-Theme scale initiates the phrase, "IN COMPARISON WITH OTHER MALES, YOUR SCORE IS MODERATELY LOWER THAN AVERAGE." However, the same score for a female would initiate the phrase "IN COMPARISON WITH OTHER FEMALES, YOUR SCORE IS ABOUT AVERAGE."

Insert Table 2 About Here

To reiterate, the person's numerical score is based on the combined male and female general reference sample, while the descriptive phrases above are based on the score distribution for the respondent's sex.

Figure 4 below illustrates the combined effect of the above three parts in the analysis of the General Occupational Themes, First is the descriptive paragraph for R-THEME which contains an explanation of the scale's meaning and examples of occupational groups that are representative of the R-THEME. All reports contain this basic paragraph for each theme scale. Second is a sentence which is dependent on the individual's score relative to a general reference sample of males and females, and it gives the person an indication of the relevancy of the preceding descriptive paragraph. Third is a sentence that relates the person's score to the population distribution relative to the individual's sex. Highlighting all this is the expanded scale name down the left and right hand sides of the page.

Insert Figure 4 About Here

A lot of space is used on the interpretive profile for an analysis of the person's interests on the six theme scales. As will be discussed later, the R-I-A-S-E-C model is used to summarized the Basic Interest Scales and Occupational Scales. Thus, by emphasizing these themes in the beginning of the interpretive report, later analyses of the person's interests will be more comprehensive and interpretation will be easier for the person.

BASIC INTEREST SCALES

The second set of scores presented on the interpretive output is the Basic Interest Scales. This type of scale, developed in the late 1960s for the SVIB, was a relatively late addition in the total history of the SVIB. Scale development was based on cluster analysis -- all pair-wise item correlations were computed separately for the male and female SVIBs, based on samples representing men-in-general and women-in-general (Campbell,

TABLE 2

INTERPRETIVE STATEMENTS FOR OCCUPATIONAL THEMES BY SEX

TEXT	REALISTIC		INVESTIGATIVE		ARTISTIC	
	MALES	FEMALES	MALES	FEMALES	MALES	FEMALES
... A VERY LOW SCORE	27-38	27-31	20-35	20-32	24-30	24-38
... A LOW SCORE	39-44	32-35	36-41	33-37	31-35	39-43
... MODERATELY LOWER THAN AVERAGE	45-49	36-39	42-46	38-42	36-39	44-49
... AN AVERAGE SCORE	50-61	40-50	47-57	43-55	40-53	50-59
... MODERATELY HIGHER THAN AVERAGE	62-64	51-56	58-61	56-59	54-59	60-62
... A HIGH SCORE	65-67	57-61	62-64	60-63	60-63	63-64
... A VERY HIGH SCORE	68-75	62-75	65-69	64-69	64-67	65-67

	SOCIAL		ENTERPRISING		CONVENTIONAL	
	MALES	FEMALES	MALES	FEMALES	MALES	FEMALES
... A VERY LOW SCORE	19-32	19-37	30-35	30-36	23-36	23-36
... A LOW SCORE	33-38	38-43	36-40	37-38	37-40	37-40
... MODERATELY LOWER THAN AVERAGE	39-43	44-47	41-45	39-43	41-44	41-45
... AN AVERAGE SCORE	44-53	48-54	46-57	44-51	45-54	46-53
... MODERATELY HIGHER THAN AVERAGE	54-60	55-61	58-63	52-56	55-58	54-60
... A HIGH SCORE	61-65	62-64	64-69	57-61	59-65	61-67
... A VERY HIGH SCORE	66-73	65-73	70-77	62-77	66-79	68-79

FIGURE 4

EXAMPLE OF R-THEME OUTPUT

```
R      R-THEME---DESCRIPTIVE EXAMPLES ARE RUGGED, PRACTICAL, AND PHYSICALLY STRONG.  SUCH PEOPLE    R
E  USUALLY HAVE TROUBLE EXPRESSING THEMSELVES IN WORDS OR COMMUNICATING THEIR FEELINGS TO OTHERS.  E
A  THEY LIKE TO WORK OUTDOORS, AND THEY LIKE TO WORK WITH TOOLS, ESPECIALLY LARGE MACHINES.  THEY   A
L  PREFER TO DEAL WITH OBJECTS RATHER THAN WITH IDEAS OR WITH PEOPLE.  THEY ENJOY CREATING THINGS   L
I  WITH THEIR HANDS AND PREFER OCCUPATIONS SUCH AS MECHANIC, CONSTRUCTION WORK, FISH AND WILDLIFE   I
S  MANAGEMENT, LABORATORY TECHNICIAN, SOME ENGINEERING SPECIALTIES, SOME MILITARY JOBS, AGRICULTURE S
T  OR THE SKILLED TRADES.  ALTHOUGH NO SINGLE WORD CAN BEST CAPTURE THE BROAD MEANING OF THE ENTIRE T
I  THEME, THE WORD REALISTIC HAS BEEN USED, THUS R-THEME.                                           I
C      YOUR SCORE OF 77 ON THE R-THEME SCALE INDICATES THAT PROBABLY MOST OF THE ABOVE DESCRIP-     C
   TIONS ARE TRUE FOR YOU AND YOU WOULD ENJOY ACTIVITIES THAT ALLOW YOU TO WORK WITH YOUR HANDS.
   IN COMPARISON WITH OTHER MALES, YOUR SCORE IS A VERY HIGH SCORE.
```

et.al. 1968; Campbell, 1971). Figure 5 illustrates a very high item intercorrelation between "Algebra" and "Arithmetic" (r=.63).

Insert Figure 5 About Here

For example, the items "Military Drill," "Military Officer," "Drilling Soldiers," "Drilling in a military company," and "Military Officers" all had high intercorrelations and all were related to the same general activity. Together, these items form the content of the Military Activities scale (Campbell, 1971).

To construct Basic Interest Scales for the SCII, male and female SVIB sets were merged into a single form. The resulting 23 SCII Basic Interest Scales were normed on the general reference sample (N=600). The standardization procedure produced a mean of 50 and standard deviation of 10 for each scale; however, there were some significant differences between male and female means, and these are noted below.

Table 3 contains basic information used in presentation of results for the Basic Interest Scales on the interpretive report; for example, the first column contains the Theme code (R,I,A,S,E,C) that is most closely related to each of the basic scales. Determination of theme codes was based on correlations with the basic scales. For example, Music/Dramatics, Art and Writing all correlate most highly with the A-Theme scale and thus are coded "A" [Campbell (1974) presents correlation data used in determination of code types.].

Insert Table 3 About Here

Next, in Table 3, are the scale names as they appear on the interpretive report, as well as the number of items that are contained within each scale. Length of the basic scales ranges from only 5 items for Military Activities to 24 items on Mechanical Activities. They generally are shorter than either the 20-item General Theme Scales or the much longer Occupational Scales (roughly 40-60 items). Although these scales are shorter, they are internally consistent and do represent an important set of scales for ana-analyzing the individual's interest.

Next are the means and standard deviations for females and males of the general sample and mean differences between the sexes. As with the theme scales, some basic scales showed substantial sex differences, one-half standard deviation and greater for many of the scales; this was especially noticeable for DOMESTIC ARTS and MECHANICAL ACTIVITIES. As with the theme scales, the interpretive report takes gender into consider-

FIGURE 5

EXAMPLE OF ITEM INTERCORRELATION

		"ALGEBRA"				
		Like	Indifferent	Dislike	Total	
	Like	51%	12%	6%	69%	
"Arithmetic"	Indifferent	3%	11%	6%	19%	r=.63
	Dislike	1%	1%	10%	12%	
	Total	55%	24%	22%	100%	

TABLE 3

MEANS, STANDARD DEVIATIONS, AND MAXIMUM AND MINIMUM POSSIBLE SCORES ON THE BASIC INTEREST SCALES FOR THE MALE AND FEMALE GENERAL REFERENCE SAMPLES

General theme	Scale	Number of Items	Female sample (N=300)		Male sample (N=300)		Mean difference, male from female	Minimum score	Maximum score
			Mean	S.D.	Mean	S.D.			
R	AGRICULTURE	7	48.59	9.98	51.41	9.98	-2.82	30	68
R	NATURE	11	51.63	9.28	48.37	10.64	3.26	17	67
R	ADVENTURE	9	46.06	10.19	53.94	9.78	-7.88	30	71
R	MILITARY ACTIVITIES	5	47.32	8.10	52.68	11.56	-5.36	41	76
R	MECHANICAL ACTIVITIES	24	45.46	10.20	54.54	9.76	-9.04	30	72
I	SCIENCE	17	48.38	10.27	51.62	9.69	-3.24	28	69
I	MATHEMATICS	9	48.27	10.14	51.73	9.83	-3.46	31	67
I	MEDICAL SCIENCE	8	49.94	10.39	50.06	9.56	-0.12	29	68
I	MEDICAL SERVICE	9	52.61	11.44	47.39	8.28	5.22	33	76
A	MUSIC/DRAMATICS	14	54.03	9.67	45.97	10.29	8.06	26	73
A	ART	14	54.59	9.31	45.41	10.61	9.18	26	68
A	WRITING	14	52.49	9.36	47.51	10.57	4.98	26	66
S	TEACHING	9	51.90	9.03	48.10	10.85	3.80	23	67
S	SOCIAL SERVICE	11	52.50	9.30	47.50	10.70	5.00	24	69
S	ATHELETICS	12	47.09	9.56	52.91	10.39	-5.82	27	70
S	DOMESTIC ARTS	10	57.30	10.40	42.70	9.55	14.60	27	75
S	RELIGIOUS ACTIVITIES	9	51.49	9.66	48.51	10.30	2.98	31	68
E	PUBLIC SPEAKING	10	47.99	9.98	52.01	9.99	-4.02	31	71
E	LAW/POLITICS	12	47.51	10.30	52.49	9.66	-4.98	30	69
E	MERCHANDISING	13	49.74	9.37	50.26	10.56	-0.52	31	72
E	SALES	12	47.30	8.40	52.70	11.40	5.40	37	81
E	BUSINESS MANAGEMENT	14	48.05	9.40	51.95	10.53	-3.90	29	72
C	OFFICE PRACTICES	16	52.89	10.82	47.11	9.07	5.78	36	79

R = Realistic; I = Investigative; A = Artistic; S = Social; E = Enterprising; C = Conventional.

ation. Both the person's standard score and the middle range of scores for males and females, depending on the individual's sex, are graphically displayed. The middle range is the mean, plus and minus two-thirds of a standard deviation, computed separately for males and females. This range of scores is presented in the basic scale profile portion of the interpretive report and corresponds roughly to the 25th through 75th percentile or middle 50 percent of the scores.

The last two columns in Table 3, minimum and maximum scores, are presented on the interpretive report to indicate the highest and lowest possible score for each scale. Counselors have indicated that such information is very useful; for example, a score of 66 on the Writing scale indicates that the maximum scores has been obtained while a score of 81 on the Sales scale is the maximum possible.

Basic Scale Profile

Figure 6 is an abbreviated example of the basic scale profile portion and indicates how the data in Table 3 have been implemented. (Page 5 of the report in Appendix 1 illustrates the profile pattern in total.)

Insert Figure 6 About Here

As indicated in Figure 6, all graphical representations of the 23 basic scales are preceded by the introductory paragraph which gives a brief explanation of the scores. The individual also is informed that adults average between 43 and 57 (again the middle 50 percent of scores for the total general reference sample), that the series of periods on the profile are the middle range of scores corresponding to the sex of the individual, and that the highest and lowest possible scores are indicated by MAX and MIN.

The order of these scores is from highest to lowest. Within this arrangement, scores are grouped into five categories: VERY HIGH, HIGH, AVERAGE, LOW, and VERY LOW. As with the general theme scales, the profile indicates at a single glance the relative strengths and weaknesses across a number of scales. Addition of the series of periods that display the average range of scores relevant to the person's sex and minimum and maximum values complicate the visual effect somewhat but after trying various alternatives, this proved to be the most effective for conveying this information without unduly cluttering the display.

Thus, the fifth page of every report will contain the individual's scores on 23 Basic Interest Scales. The following characterized this section of the individualized report:

1. Scores on the 23 basic scales are ranked from highest to lowest.

FIGURE 6

ABBREVIATED EXAMPLE OF BASIC INTEREST SCALE PROFILE

BELOW ARE YOUR RESULTS ON 23 BASIC INTEREST SCALES. THEY SHOW THE STRENGTH OF YOUR INTERESTS IN A VARIETY OF PURE TYPES OF AREAS. THESE SCORES WILL SHOW AREAS WHERE YOU WILL FIND SATISFACTION OR DISSATISFACTION. THEY CAN BE IMPORTANT IN MAKING AN OCCUPATIONAL CHOICE OR IN DECIDING ON LEISURE-TIME ACTIVITIES. THE AVERAGE ADULT SCORES BETWEEN 43 AND 57 ON EACH SCALE. HIGH SCORES SHOW AREAS WHERE YOU WILL FIND ENJOYABLE TYPES OF ACTIVITIES. LOW SCORES INDICATE AREAS THAT WILL BE LESS REWARDING TO YOU. ON SOME SCALES MALES AND FEMALES RESPOND SOMEWHAT DIFFERENTLY AND THE RANGE OF DOTS INDICATES THE AVERAGE RANGE OF SCORES FOR YOUR SEX. ALSO, THE LOWEST POSSIBLE SCORE (MIN) AND HIGHEST POSSIBLE SCORE (MAX) ARE LISTED.

```
                         YOUR        3         4         5         6         7
BASIC INTEREST SCALES   SCORE    ....0.........0.........0.........0.........0....     MIN-MAX

****VERY HIGH (66+)****

****HIGH (65-58)****
WRITING                  65                    ...............           *             26-66
PUBLIC SPEAKING          64                       ..................    *              31-71
LAW/POLITICS             63                        .................   *               30-69
MATHEMATICS              61                        ................  *                 31-67
MUSIC/DRAMA              58                 ..................    *                    26-73

****AVERAGE (57-43)****
SCIENCE                  57                        .............*...                   28-69
NATURE                   54                      ............*....                     17-67
```

FIGURE 7

EXAMPLE OF HIGH SCORE ANALYSIS ON BASIC INTEREST SCALE

YOUR HIGHEST SCORES ON THE BASIC INTEREST SCALES ARE LISTED BELOW WITH EXAMPLES OF OCCUPATIONS THAT SCORE HIGH AND LOW. PEOPLE IN OCCUPATIONS THAT ARE EXAMPLES OF HIGH SCORES LIKE THE ACTIVITIES THAT ARE MEASURED BY THAT SCALE. PEOPLE IN OCCUPATIONS THAT ARE EXAMPLES OF THE LOW SCORES HAVE STRONG DISLIKES FOR THOSE SAME ACTIVITIES. THE HIGHER YOUR SCORE, THE MORE SIMILAR YOUR INTERESTS ARE TO PEOPLE IN THE HIGH SCORING OCCUPATIONS AND THE MORE DISSIMILAR YOUR INTERESTS ARE TO PEOPLE IN OCCUPATIONS THAT SCORE LOW.

ALSO, PAGE REFERENCES ARE PRESENTED FOR THE OCCUPATIONAL OUTLOOK HANDBOOK (OOH), BULLETIN NO. 1785 THAT CAN BE USED FOR FURTHER REFERENCE. THE OOH IS A GOOD SOURCE FOR RECENT TRENDS IN EMPLOYMENT OPPORTUNITIES AND EDUCATIONAL REQUIREMENTS FOR VARIOUS OCCUPATIONS THAT ARE RELATED TO YOUR HIGH AND LOW SCORES. THE OOH CAN BE FOUND IN YOUR LOCAL LIBRARY OR ORDERED DIRECTLY FROM THE U.S. GOVERNMENT PRINTING OFFICE, WASHINGTON, D.C., 20402 ($6.85).

YOUR ---HIGHEST SCORES---	---EXAMPLES OF HIGH SCORES-----	-----EXAMPLES OF LOW SCORES-----
69 WRITING OOH 606-615	ADVERTISING EXECUTIVE, WRITER, MINISTER, REPORTER	FACTORY WORKER, CARPENTER, BEAUTICIAN, FARMER
64 PUBLIC SPEAKING OOH 141-151	SALES PERSONNEL, MINISTER, STATE LEGISLATOR, LAWYER	BEAUTICIAN, CARPENTER, ARTIST, ARCHITECT, FARMER
63 LAW/POLITICS OOH 141-144	POLITICAL SCIENTIST, LAWYER JUDGE, STATE GOVERNOR	ARCHITECT, BEAUTICIAN, ARTIST, FARMER, PHYSICIST

YOUR HIGHEST SCORES ON THE BASIC INTEREST SCALES INDICATE YOUR STRONGEST PREFERENCE FOR THE PURE TYPES OF INTERESTS AND ACTIVITIES MEASURED BY THESE SCALES. IN GENERAL, THESE AREAS PROBABLY WILL BE SOURCES OF SATISFACTION IN YOUR LIFE. ALTHOUGH THEY MAY OR MAY NOT BE PART OF YOUR OCCUPATIONAL CHOICE, YOU PROBABLY WILL FIND YOUR OCCUPATIONAL CHOICE MORE REWARDING IF THE INTERESTS ARE INCLUDED. YOUR HIGH SCORES ALSO WILL POINT OUT AREAS OF SATISFACTION FOR LEISURE TIME ACTIVITIES.

FIGURE 8

INTERPRETATION OF HIGH SCORES ON BASIC INTEREST SCALE

TO GIVE YOU MORE INFORMATION ABOUT YOUR HIGH SCORES, THE FOLLOWING GENERAL STATEMENTS WILL SUMMARIZE THE IMPORTANT PARTS OF THESE SCORES. GENERALLY, THE HIGHER YOUR SCORE THE MORE APPLICABLE THE PHRASES ARE FOR YOU. YOU HAVE INDICATED THAT YOU LIKE--

---AREAS DEALING WITH WRITING AND CREATIVITY IN EXPRESSING THOUGHTS THROUGH WORDS.
---SPEAKING IN FRONT OF PEOPLE AND TALKING TO LARGE GROUPS.
---OCCUPATIONS THAT DEAL WITH LEGAL MATTERS AND POLITICAL DISCUSSIONS.

2. Scores are presented under five descriptive headings that relate the individual's score to the score distribution of a general reference sample.
3. Each score is plotted graphically on a standard score scale and represented by a single asterisk.
4. For each scale, the middle 50 percent of scores in the general reference sample for the individual's sex is represented by a series of periods.
5. The highest and lowest possible scores for each scale are listed.

Basic Scale Analysis

Following the one-page profile of the Basic Interest Scales, the score meanings are explained in greater depth and cross-references to specific pages in the Occupational Outlook Handbook (OOH). Figure 7 illustrates the first portion of this analysis. First is a general discussion of the scores and an explanation of the OOH.

Insert Figure 7 About Here

Following this are the three highest scores (plus ties), page references to the OOH, and examples of high and low scoring occupations for each of the listed scales. Archive files at the University of Minnesota permitted analysis of basic interest scores for over 300 occupational samples that ranges in size from 13 to over 1000 individuals. Examples of occupations that score high and low on each scale are drawn from Campbell (1974).

Analysis of the highest scores continues by stressing their import; that is, that they mainly are avocational measures of interest and the activities and interests measured by these scales will be a source of satisfaction to the person even though they may or may not be a part of an occupational choice. To further the understanding of the scores, the interpretive report presents a short descriptive summary of the item composition for the scale. As indicated in Figure 8, one-line sentences are used to convey the meaning of each of the basic interest scales that appeared above.

Insert Figure 8 About Here

Figure 9 illustrates the analysis of low scores; the three lowest scores (plus ties) on the basic scales are listed, the occupational examples are reversed from the previous score presentation so that low scoring occupations are listed first, followed by examples of high scoring occupations and relevant OOH pages. As with the high scores, a brief explanatory sentence

FIGURE 9

EXAMPLE OF LOW SCORE ANALYSIS ON BASIC INTEREST SCALES

THE FOLLOWING ARE YOUR LOWEST SCORES ON THE BASIC INTEREST SCALES. THEY INDICATE AREAS WHERE YOUR INTEREST PREFERENCES ARE LEAST STRONG FOR THE ACTIVITIES AND AREAS MEASURED BY THESE SCALES. YOU PROBABLY WILL HAVE LESS SATISFACTION WITH A CAREER RELATED DIRECTLY TO YOUR LOWEST SCORES THAT A CAREER RELATED TO YOUR HIGHEST SCORES. YOUR LOWEST SCORES ARE MOST SIMILAR TO THE LOW SCORING OCCUPATIONS AND LEAST SIMILAR TO THE OCCUPATIONS THAT SCORE HIGH.

YOUR ---LOWEST SCORES---	-----EXAMPLES OF LOW SCORES-----	----EXAMPLES OF HIGH SCORES-----
26 MECHANICAL OOH 245-284	LANGUAGE TEACHER, LAWYER, SALES PERSONNEL, POLITICAL SCIENTIST	SKILLED TRADES, ENGINEER, DENTIST CARPENTER, MACHINIST, CHEMIST
37 MILITARY ACTIVS. OOH 799-800	PSYCHOLOGIST, SOCIOLOGIST, ACTOR, MUSICIAN, ANTHROPOLOGIST	MILITARY OFFICER, POLICE OFFICER, AIRPLANE PILOT
39 ADVENTURE OOH 179-189	BANKER, JUDGE, SCHOOL SUPERINTENDENT, MATHEMATICIAN	ASTRONAUT, POLICE OFFICER, SALES PERSONNEL, MILITARY OFFICER

YOUR LOWEST SCORES ON THE BASIC INTEREST SCALE INDICATE YOUR WEAKEST PREFERENCE FOR THE PURE TYPES OF INTERESTS AND ACTIVITIES MEASURED BY THESE SCALES. IN GENERAL, YOU PROBABLY WILL FIND OTHER AREAS MORE SATISFYING THAN THOSE INDICATED BY YOUR LOW SCORES.

TO GIVE YOU MORE INFORMATION ABOUT THESE SCORES, THE FOLLOWING GENERAL STATEMENTS WILL SUMMARIZE THE IMPORTANT ASPECTS: GENERALLY, THE LOWER YOUR SCORE THE MORE APPLICABLE THE STATEMENT IS FOR YOU. YOUR ANSWERS TO THE INVENTORY HAVE INDICATED THAT YOU DO NOT LIKE--

---OPERATING MACHINES, TOOLS, AND GENERALLY WORKING WITH YOUR HANDS.
---ACTIVITIES AND PEOPLE ASSOCIATED WITH REGIMENTATION SUCH AS MILITARY LIFE.
---DANGEROUS TYPES OF ACTIVITIES. YOUNG ADULTS SCORE HIGH BUT SCORE LOWER WITH INCREASING AGE.

FIGURE 10

EXAMPLE OF BASIC INTEREST SCALE SUMMARY IN RELATION TO THEME SCALES

THE PRECEDING HAS BEEN AN ANALYSIS OF YOUR INTERESTS ON 23 BASIC INTEREST SCALES. THESE SCALES ALSO ARE RELATED TO THE PREVIOUS 6 GENERAL OCCUPATIONAL THEME SCALES. SUMMARIZED BELOW ARE YOUR SCORES ON EACH THEME SCALE AND THE RELATED BASIC INTEREST SCALES. REMEMBER, IN A GENERAL POPULATION, THE AVERAGE ADULT HAS A SCORE OF 50 ON EACH SCALE. CONSISTENTLY HIGH OR LOW SCORES IN AN AREA SHOULD BE IMPORTANT INDICATORS FOR YOU IN UNDERSTANDING YOUR INTERESTS.

I. 77 R-THEME ---
(REALISTIC)

75 AGRICULTURE
72 NATURE
66 ADVENTURE
55 MILITARY ACTIVITIES
55 MECHANICAL

II. 70 I-THEME ---
(INVESTIGATIVE)

67 SCIENCE
65 MATHEMATICS
64 MEDICAL SERVICE
54 MEDICAL SCIENCE

III. 50 A-THEME
(ARTISTIC)

60 MUSIC/DRAMA
59 ART
57 WRITING

IV. 40 S-THEME ---
(SOCIAL)

50 SOCIAL SERVICE
43 TEACHING
42 ATHLETICS
40 DOMESTIC ARTS
39 RELIGIOUS ACTIVITIES

V. 40 E-THEME ---
(ENTERPRISING)

60 PUBLIC SPEAKING
59 LAW/POLITICS
57 MERCHANDISING
55 SALES
50 BUSINESS MANAGEMENT

VI. 30 C-THEME ---
(CONVENTIONAL)

35 OFFICE PRACTICE

is elicited for each of the 3 lowest scores (plus ties) that are less that 44.

Insert Figure 9 About Here

Thus, the sixth and seventh pages of the interpretive report will contain an analysis of the individual's three highest and three lowest basic interest scales. Page references to the Occupational Outlook Handbook are presented, examples of occupational samples that score high and low on each scale are given, and a cursory descriptive sentence about the item content of those scales greater than 56 or less than 44 are printed.

Basic Scale Summary

The final analysis on the Basic Interest Scales is a summary section that ties these scales to the Occupational Theme Scales. As illustrated in Figure 10, scores on the six theme scales are listed as headings for the relevant basic scales and their corresponding scores. Theme scales are presented from highest to lowest across the page, headed by Roman numerals I-VI, while the basic scales are rank ordered from highest to lowest down the page.

Insert Figure 10 About Here

OCCUPATIONAL SCALES

Empirically constructed occupational scales on the SVIBs have been the forte of the inventory and have represented the most specific level of aggregated measurement of a person's interest preferences. The historically heavy emphasis on these scales has been carried through the latest revision, and the current SCII contains 124 occupational scales.

Occupational scales are based on item response percentage differences between the occupation and a general reference sample. The larger percentage difference, the more discriminating the item is in separating a criterion sample from an "average" sample. For those differentiating items that had significant differences between the criterion sample and reference sample, the "Like" or "Dislike" responses were weighted +1 or -1 in favor of the criterion sample. For example, Table 4 exemplifies the weighted responses for items that comprise the female Lawyer scale. Lawyers respond "Like" more frequently than the reference sample to the item "Corporation Lawyer" and thus the "Like" response is weighted +1. In contrast, lawyers respond "Dislike" more frequently than the reference sample to the item "Art Teacher" and thus the "Dislike" response is weighted -1. High scores on the female Lawyer scale are achieved by a combination of "Like" responses to those items listed in the upper half of Table 4 and "Dislike" to those items in the lower one-half of the table.

Insert Table 4 About Here

After scale development, individuals in the criterion sample are used to norm the scale. Through the raw-score-to-standard-score transformation procedure, each criterion group has a mean of 50 and standard deviation of 10 on their own scale. People not in the occupation usually average about 30, or about two standard deviations below the criterion mean.

One modification of the SVIB reporting schema is the use of descriptive headings for occupational scores on the SCII. The traditional grade-letter headings have been eliminated in favor of descriptive headings such as SIMILAR and DISSIMILAR. These headings relate the magnitude of the individual's score to those of people in the criterion sample. The use of these new headings will help individuals think more in terms of similarities and not in terms of ability rankings as was sometimes the case with the historical letter grades.

Just as response differences between males and females were a concern in reporting results for the General Occupational Scales and Basic Interest Scales, so too were response differences a concern for the Occupational Scales. Since occupational scales are based on percentage differences, combination of male and female responses would have decreased the validity of the scales. Thus, occupational scales were developed separately for males and females. For example, a criterion sample of female artists was used to develop a female Artist scale and a criterion sample of male artists was used to develop the male Artist scale.

To provide as much information as possible, interpretive reports for the SCII Occupational scales consist of three sections:

1. The first section is a presentation of scores on the occupational scales that correspond to the individual's sex, that, like-sex occupational scales.
2. The second section is an analysis of the high and low score for the like-sex scales.
3. The third section is a listing of scores for the occupational scales based on the opposite sex, that is, opposite-sex occupational scales.

Occupational Scale Report -- Like-Sex Scales

The first page of the interpretive report of like-sex occupational scales begins on page 9 of each interpretive report. As indicated in

TABLE 4

EXAMPLE OF ITEMS ON LAWYERS (F) SCALE

ITEM NUMBER	ITEM WEIGHTS			ITEM NAME
	L	I	D	
35	+1	0	-1	CORPORATION LAWYER
38	+1	0	-1	CRIMINAL LAWYER
46	+1	-1	-1	EDITOR
58	+1	0	-1	GOVERNOR OF A STATE
65	+1	0	-1	INCOME TAX ACCOUNTANT
69	+1	-1	-1	JUDGE
85	+1	-1	-1	NEWSPAPER REPORTER
98	+1	0	-1	POLITICIAN
142	+1	-1	-1	CIVICS (GOVERNMENT)
144	+1	-1	-1	ECONOMICS
145	+1	-1	-1	ENGLISH COMPOSITION
149	+1	0	-1	JOURNALISM
5	-1	0	+1	ART TEACHER
18	-1	0	+1	BEAUTY AND HAIRCARE CONSULTANT
40	-1	0	+1	DENTAL ASSISTANT
45	-1	0	+1	DRESSMAKER/TAILOR
60	-1	0	+1	HOME ECONOMICS TEACHER
61	-1	0	+1	HOSPITAL RECORDS CLERK
62	-1	0	+1	HOUSEKEEPER
73	-1	0	+1	MGR, CHILD CARE CENTER
86	-1	0	+1	NURSE
95	-1	0	+1	PLAYGROUND DIRECTOR
107	-1	0	+1	RECEPTIONIST
115	-1	0	+1	SOCIAL WORKER
119	-1	0	+1	STEWARD/STEWARDESS
130	-1	0	+1	X-RAY TECHNICIAN

Figure 11, the actual scores are preceded by two introductory paragraphs, which are uniform for all reports. Like other sections of the interpretive report, a conscious effort has been made to avoid wording, phrasing, or referencing anything that may lead the individual to over-interpret the results in terms of either ability or success. Although minor changes may be made over the years in wording or phrasing, the essential meaning will remain about the same.

The like-sex occupational scales are ranked from highest to lowest and are presented under five descriptive headings: "VERY SIMILAR," "SIMILAR," "AVERAGE," "DISSIMILAR," and "VERY DISSIMILAR." These five headings indicate the extent that the individual's interests match workers of the same sex in specific occupational samples. Through use of these descriptive headings the person should think in the frame of reference about how her or his interest preferences compare with people in that occupation.

Insert Figure 11 About Here

Occupational Scale Analysis

DOT and OOH References

To focus the person's attention on occupational areas where chances are greatest for satisfaction or dissatisfaction, high and low scores on the like-sex occupational scales are analyzed further. First, there is a general introduction about satisfaction in relation to high and low scores; that is, high scores indicate occupations where the individual's likes and dislikes most closely match workers of the same sex in that occupation and vice versa for low scores. The Dictionary of Occupational Titles (DOT) also is introduced, that it contains information about duties and functions of each occupation, and that is can be ordered from the U.S. Government Printing Office.

Next there is a short sentence preceding both the high and low scores about similarity or dissimilarity to employed males or females in the occupations. Results for males make reference to "MALES" as indicated in Figure 12, and reports for females make reference to "FEMALES." These generic references were included explicitly to insure that the person is aware that scores being reported do not refer to a general sample of workers in the occupation, but rather to workers categorized by sex.

After these sentences, the four highest like-sex occupational scales (plus ties) and the four lowest like-sex occupational scales (plus ties) are presented. References to the DOT occupational coding system are listed and specific page references for the Occupational Outlook Handbook (OOH) also are given.

FIGURE 11

EXAMPLE OF LIKE-SEX OCCUPATIONAL SCALES (MALE)

THE NEXT GROUP OF RESULTS ARE OCCUPATIONAL SCALES. THEY INDICATE THE DEGREE OF SIMILARITY BETWEEN YOUR INTERESTS AND THOSE OF EMPLOYED PEOPLE WHO ARE OF THE SAME SEX AS YOU. THE AVERAGE WORKER IN AN OCCUPATION SCORES ABOUT 50 ON THAT SCALE AND ABOUT 2/3 OF THESE WORKERS HAVE SCORES BETWEEN 40 and 60. MOST PEOPLE NOT IN THE OCCUPATION SCORE ABOUT 30 ON THE SCALE.

RESEARCH HAS INDICATED THAT PEOPLE WHO ENTER AN OCCUPATION WHERE THEY HAVE HIGH SCORES TEND TO REMAIN IN THAT OCCUPATION AND ARE MORE SATISFIED THAN IF THEY ENTER AN OCCUPATION WHERE THEY HAVE LOW SCORES. FURTHERMORE, THESE SCORES MAY INDICATE INTEREST IN AN OCCUPATION THAT YOU MAY NOT HAVE CONSIDERED BEFORE AND CAN HELP YOU THINK ABOUT OTHER VARIOUS CAREER POSSIBILITIES.

VERY SIMILAR
55+

55 PUBLIC ADMINIS.
55 SCHOOL SUPERDT.

SIMILAR
54-45

54 PERSONNEL DIR.
54 SOCIAL WORKER
52 PSYCHOLOGIST
51 PSYCHIATRIST
49 LAWYER
46 LIBRARIAN
46 CH. OF COMM. EXEC.
46 REGISTERED NURSE
46 PHYS. SCIENTIST
45 COLLEGE PROF.
45 SPEECH PATHOL.
45 PHYS. THERAPIST

AVERAGE
44-26

44 CARTOGRAPHER
44 MEDICAL TECH.
42 LIFE INS. AGENT
42 ADVERTISING
40 MATHEMATICIAN
38 BIOLOGIST
38 REALTOR
34 MINISTER
34 COMPUTER PROGRM.
32 SALES MANAGER
31 CHEMIST
31 MUSICIAN
31 MUSIC TEACHER
31 CREDIT MANAGER
31 RECREATION LEADER
30 OPTOMETRIST
28 SENIOR C.P.A.
28 ACCOUNTANT
27 YMCA STAFF
27 PHYSICIAN
27 PHYSICIST
26 SOC. SCI. TEACH.
26 BUS. ED. TEACHER
26 SOCIAL SCIENTIST
26 REPORTER

DISSIMILAR
25-16

25 ARTIST
24 ENGINEER
24 PRODUCTION MGR.
23 OSTEOPATH
23 PHYSICAL THERP.
22 AIR FORCE OFF.
22 PURCHASG. AGENT
21 ARCHITECT
21 OFFICE WORKER
21 BANKER
21 FUNERAL DIRCT.
19 MATH-SCI. TEACH.
19 PHARMACIST
18 PHOTOGRAPHER
18 INT. DECORATOR
18 ADVERTISING EXEC.
18 ENGLISH TEACHER
17 LIC. PRAC. NURSE

VERY DISSIMILAR
15-

15 DENTIST
15 ARMY OFFICER
15 FARMER
11 VETERINARIAN
10 NAVY OFFICER
10 MERCH. MARINE OFF.
10 DIETITIAN
9 PRIEST
5 HWY. PATROL OFF.
3 FORESTER
3 POLICE OFFICER
-8 SKILLED CRAFTS
-8 VOC. AGRIC. TCHR.

Insert Figure 12 About Here

OOH Reprints Series

Following references to the DOT and OOH, the high scores that were presented in the previouw section are keyed into additional data files. Figure 13 illustrates the next sequence of the interpretive report. Reprints from the OOH are listed that are most relevant to each specific occupational scale. Included are the bulletin number, cost of the reprint, and reprint title needed to order from the U.S. Government Printing Office. The reprint pamphlets are actually subsections of the complete Occupational Outlook Handbook and, thus, they are redundant if access to the OOH is possible.

Insert Figure 13 About Here

In some instances, more than one occupational scale is keyed to the same reprint bulletin; for example, all the military scales such as Air Force Officer, Navy Officer, and Army Officer are related to the government service reprint. To avoid redundancy, only unique entries are listed. In a few situations, such as School Superintendent, no reprint reference was deemed relevant; consequently the number of references that are listed in this section for some individuals will be small; for example, a person may have four or five high occupational scales under consideration, but only two or three reprint references may be listed.

Book References

Besides references to the reprint series, book references are presented as additional sources of information for exploring the individual's interests. Figure 14 illustrates this feature. For most scales, a book reference was found to be useful and helpful and was included in the data file for the occupation. However, for some occupations (about 10%) a relevant book was not found. These situations, and cases where more than one occupational scale references the same book, will influence the number of additional references that are presented since only unique, non-blank data file entries are listed.

Insert Figure 14 About Here

Activities

Following the additional book references, a series of suggested activities

FIGURE 12

EXAMPLE OF OCCUPATIONAL SCALE ANALYSIS
HIGH AND LOW SCORES

BELOW ARE YOUR HIGHEST AND LOWEST OCCUPATIONAL SCORES. YOU SHOULD PAY PARTICULAR ATTENTION TO THOSE OCCUPATIONS WHERE YOUR SCORES INDICATED THAT YOU HAD THE MOST SIMILAR INTERESTS AND TO THOSE OCCUPATIONS WHERE YOUR INTERESTS WERE MOST DISSIMILAR. YOU WILL HAVE THE BEST CHANCE OF FINDING SATISFACTION IF YOU DECIDE ON AN OCCUPATION WHERE YOUR INTERESTS ARE SIMILAR WITH YOUR CO-WORKERS AND LESS CHANCE IF YOUR INTERESTS ARE DISSIMILAR. IF YOU RECORDED THE SAME LIKES AND DISLIKES AS THE WORKERS, YOUR SCORE WILL BE HIGH FOR THAT OCCUPATION. IF YOUR LIKE AND DISLIKE RESPONSES WERE DIFFERENT FROM THOSE OF PEOPLE IN THE OCCUPATION, YOUR SCORE WILL BE LOW AND YOU WOULD NOT LIKELY BE HAPPY IN THAT KIND OF WORK.

AS BEFORE, REFERENCES ARE GIVEN FOR THE OCCUPATIONAL OUTLOOK HANDBOOK (OOH) WHICH PRESENTS ADDITIONAL INFORMATION ON EMPLOYMENT OPPORTUNITIES AND RELEVANT WORK SITUATIONS. ALSO, REFERENCES ARE LISTED FOR THE DICTIONARY OF OCCUPATIONAL TITLES (DOT) VOLUME I, THIRD EDITION. THE DOT GIVES DETAILED JOB DESCRIPTIONS OF THE DUTIES AND FUNCTIONS OF EACH OCCUPATION AND CAN BE FOUND IN YOUR LOCAL LIBRARY OR PURCHASED FROM THE U.S. GOVERNMENT PRINTING OFFICE ($7.75).

YOUR HIGHEST SCORES APPEARED ON THE FOLLOWING SCALES AND INDICATE THE GREATEST DEGREE OF SIMILARITY BETWEEN YOUR ANSWERS AND THOSE OF MALES IN THESE OCCUPATIONS--

	55 PUBLIC ADMINISTRATOR	55 SCHOOL SUPERINTENDENT	54 PERSONNEL DIRECTOR	54 SOCIAL WORKER
DOT	188.118	091.118	166.118	195.108
OOH	132-134	NO REFERENCE	146-147	570-573

YOUR LOWEST SCORES APPEARED ON THE FOLLOWING SCALES AND INDICATE THE GREATEST DISSIMILARITY BETWEEN YOUR ANSWERS AND THOSE OF MALES IN THESE OCCUPATIONS--

	-8 VOC. AGRIC. TCHR.	-8 SKILLED CRAFTS	3 POLICE OFFICER	3 FORESTER
DOT	091.228	860.281-865.887	375.118-.268	040.081
OOH	625	245-284	185-187	332-333

FIGURE 13.

EXAMPLE OF OCCUPATIONAL SCALE ANALYSIS
OCCUPATIONAL OUTLOOK REPRINT SERIES

THE U.S. GOVERNMENT PRINTING OFFICE ALSO HAS A REPRINT SERIES THAT WILL PROVIDE YOU WITH ADDITIONAL INFORMATION ON EMPLOYMENT, WORKING CONDITION, AND PAY FOR MANY OCCUPATIONS. REPRINT BULLETINS ARE SUBSECTIONS OF THE OOH IN PAMPHLET FORM. LISTED BELOW IS THE BULLETIN NUMBER AND COST FOR EACH REPRINT THAT WOULD BE IMPORTANT FOR YOU IF THE OOH IS NOT READILY AVAILABLE.

BULLETIN NO.	COST	TITLE
26	$.25	CITY MANAGERS
23	$.25	PERSONNEL WORKERS
107	$.25	SOCIAL WORKERS, SOCIAL SERVICE AIDES

FIGURE 14

EXAMPLE OF OCCUPATIONAL SCALE ANALYSIS
BOOK REFERENCES RELATED TO HIGH SCORE

YOUR HIGH OCCUPATIONAL SCALE SCORES SHOULD BE VIEWED AS IMPORTANT INDICATORS OF HOW YOUR INTERESTS MATCH THOSE OF EMPLOYED WORKERS. YOU SHOULD USE THESE RESULTS AS GUIDES AND SEARCH OUT ADDITIONAL INFORMATION ABOUT THOSE AREAS WHERE YOUR INTERESTS AND ABILITIES ARE FOCUSED. ASK YOUR LIBRARIAN FOR FURTHER INFORMATION ON JOBS IN THESE AREAS, TALK TO PEOPLE IN THESE FIELDS, READ BOOKS AND PERIODICALS THAT ARE RELEVANT, AND SO FORTH.

TO LEARN MORE ABOUT THE LIFE AND WORK OF PEOPLE IN OCCUPATIONS WHERE YOU HAVE HIGH SCORES, THE FOLLOWING BOOKS ARE SUGGESTED FOR FURTHER INFORMATION--

--COHN, ANGELO. CAREERS IN PUBLIC PLANNING AND ADMINISTRATION. WALCK, 1966.
--SULLIVAN, NEIL VINCENT. WALK, RUN, OR RETREAT. INDIANA UNIVERSITY PRESS, 1971.
--FRENCH, WENDELL L. THE PERSONNEL MANAGEMENT PROCESS. HOUGHTON, 1970.
--PERLMAN, HELEN HARRIS. SO YOU WANT TO BE A SOCIAL WORKER. HARPER, 1970.

are listed corresponding to the person's four highest scores (see Figure 15). Generally, an activity is indicated that is available to most people. For a few occupational scales, the suggested activity may be more difficult than for other scales. For example, a high score on the Merchant Marine Officer scale elicits the suggestion to visit a harbor or take an ocean boat ride -- rather difficult for someone living in the land-locked states of the midwest. However, most suggestions are similar to the one for the Physician scale, where a high score suggests visiting a hospital or clinic.

Insert Figure 15 About Here

Occupational Scale Summary

After the activities section, a summary is done of the analysis that has occurred on the three types of scales (theme, basic and occupational scales). Just as the Basic Interest Scales were summarized in relation to the General Occupational Theme scales, the Occupational Scales are summarized via the theme scales.

The theoretical classification system of Holland was extended to the occupational scales and the six themes were used to organize the 124 occupational scales (Campbell, 1974). Basically, each occupation was assigned code types corresponding to R,I,A,S,E, or C on the basis of: 1) group means on the six theme scales; 2) the correlations between the occupational scale and the theme scales; 3) correlations between the occupational scale and the basic scales, and 4) correlations between the occupational scales themselves.

As indicated in the two introductory paragraphs in Figure 16, the person is reminded of the average score range for each of the three scale types (General Occupational Themes, Basic Interest Scales, and Occupational Scales). Presentation of all scores is from highest to lowest. First are the six theme scales, high to low from left to right, set off by Roman numerals I-VI. Second are the 23 related Basic Interest Scales within each theme area. Third are the four highest (plus ties) like-sex occupational scales within each theme area.

Insert Figure 16 About Here

Occupational Scale Report -- Opposite Sex Scales

The preceding sections involved analysis and presentation of like-sex occupational scales. The next portion of the interpretive results deals with the person's scores on opposite-sex occupational scales. As indicated in Figure 17 below, there is a brief introductory statement about sexual stereotypes suggesting that one's choice of a career not be limited to traditional

FIGURE 15

EXAMPLE OF OCCUPATIONAL SCALE ANALYSIS
ADDITIONAL ACTIVITIES RELATED TO HIGH SCORES

TO LEARN MORE ABOUT THE DAILY ROUTINES OF PEOPLE IN OCCUPATIONS WHERE YOU HAVE HIGH SCORES, YOU MAY WISH TO CONSIDER SPENDING SOME TIME TRYING THESE ACTIVITIES--

--VISIT THE CITY OR COUNTY ADMINISTRATIVE OFFICES OF LOCAL GOVERNMENT.
--SPEND TIME WITH A SCHOOL SUPERINTENDENT OR ATTEND A SCHOOL BOARD MEETING.
--VISIT WITH AN EMPLOYMENT AGENCY AND TALK WITH THE PERSONNEL DIRECTOR.
--DO SOME SOCIAL SERVICE VOLUNTEER WORK.

FIGURE 16

EXAMPLE OF OCCUPATIONAL SCALE SUMMARY IN REALTION TO THEME SCALES AND BASIC INTEREST SCALES

THE PREDEDING HAS BEEN AN ANALYSIS OF YOUR INTEREST ON THE OCCUPATIONAL SCALES. THESE SCALES ALSO ARE RELATED TO THE PREVIOUS 6 GENERAL OCCUPATIONAL THEME SCALES AND BASIC INTEREST SCALES. SUMMARIZED BELOW ARE YOUR SCORES ON EACH THEME SCALE, RELATED BASIC INTEREST SCALES, AND HIGHEST OCCUPATIONAL SCALES WITH EACH AREA.

REMEMBER, IN A GENERAL POPULATION, THE AVERAGE ADULT HAS A SCORE OF 50 ON EACH OF THE 6 THEME SCALES AND 23 BASIC INTEREST SCALES. HOWEVER, ON THE OCCUPATIONAL SCALES, SCORES OF 50 INDICATE SIMILARITY WITH THE AVERAGE WORKER IN THE OCCUPATION AND SCORES OF 26-44 ARE AVERAGE FOR PEOPLE-IN-GENERAL.

I. 69 I-THEME ---
(INVESTIGATIVE)

69 SCIENCE
67 MATHEMATICS
63 MEDICAL SCIENCE
50 MEDICAL SERVICE

64 PSYCHOLOGIST
62 COMPUTER PROGRAMMER
61 SPEECH PATHOLOGIST
59 SOCIAL SCIENTIST

II. 61 A-THEME ---
(ARTISTIC)

56 ART
55 WRITING
54 MUSIC/DRAMA

55 MUSICIAN
54 ARCHITECT
49 PHOTOGRAPHER
48 ARTIST

III. 56 R-THEME ---
(REALISTIC)

59 MECHANICAL
58 AGRICULTURE
58 NATURE
53 ADVENTURE
41 MILITARY ACTIVITIES

50 CARTOGRAPHER
46 VETERINARIAN
45 NAVY OFFICER
42 MERCHANT MARINE OFF.

IV. 50 C-THEME ---
(CONVENTIONAL)

44 OFFICE PRACTICES

13 BUSINESS EDUC. TCHR.
9 ACCOUNTANT
7 BANKER

V. 42 S-THEME ---
(SOCIAL)

52 TEACHING
44 DOMESTIC ARTS
43 SOCIAL SERVICE
39 ATHLETICS
33 RELIGIOUS ACTIVITIES

44 ELEMENTARY TEACHER
34 PHYSICAL THERAPIST
30 PUBLIC ADMINISTRATOR
27 LICENSED PRAC. NURSE
27 PERSONNEL DIRECTOR

VI. 38 E-THEME ---
(ENTERPRISING)

49 PUBLIC SPEAKING
46 LAW/POLITICS
41 BUSINESS MANAGEMENT
37 SALES
34 MERCHANDISING

38 INVESTMENT FUND MGR.
30 CHIROPRACTOR
27 PHARMACIST
26 LAWYER

roles (as with many other sections, space limitations on the report preclude a detailed discussion of this topic). Then either one of the two following statements appears:

> "YOUR SCORES ON SCALES DEVELOPED ON FEMALE OCCUPATIONAL SAMPLES"
> "YOUR SCORES ON SCALES DEVELOPED ON MALE OCCUPATIONAL SAMPLES"

Reports where sex has been coded as male will receive the first statement, and scores will be presented for occupational scales based on female criterion samples; reports where sex has been indicated as female will will have the second statement, followed by scores on male-based occupational samples. Figure 17 exemplifies the presentation of female-based scales for the male interpretive report. Scores are rank ordered from highest to lowest, and are presented left to right across the page.

Insert Fiugre 17 About Here

ADMINISTRATIVE SUMMARIES

LP, IP, DP

Figure 18 shows the layout of the administrative summaries and non-occupational scales for the interpretive report. Like, Indifferent, and Dislike response percentages (LP, IP, DP) to all seven sections of the SCII are included. For example, the first 131 items of the SCII are occupational titles and as indicated in the example below, the person responded "Like" to 80 percent of these items, "Indifferent" to 16 percent of these items, and "Dislike" to 4 percent.

Insert Figure 18 About Here

Total Responses (TR) and
Infrequent Responses (INF)

Total response (TR) index represents the number of responses scores and is a gross internal check on answering the SCII. If all items are answered with heavy dark marks, the total will equal the number of test items on the SCII, 325. The second internal check on answering the SCII is the infrequent response total (INF). INF indices have been developed separately for males and females; they are composed of items that have a low base rate popularity. For example, very few females respond "Like" to the item "Military Drill" while very few males respond "Like" to the item "Housekeeper." A negative score on the INF index indicates that the person has given an inordinate number of infrequent responses and may indicate that the person responded unusually which is information that the counselor may wish to pursue.

FIGURE 17

EXAMPLE OF OPPOSITE-SEX OCCUPATIONAL SCALES FOR A MALE INTERPRETIVE REPORT

THE FOLLOWING ARE YOUR SCORES ON OCCUPATIONAL SCALES DEVELOPED ON THE OPPOSITE SEX. ALTHOUGH OCCUPATIONAL SCALES DEVELOPED ON THE SEX CORRESPONDING TO YOURS ARE MORE LIKELY TO BE BETTER PREDICTORS FOR YOU THAN SCALES FOR THE OTHER SEX WOULD BE, THE RESULTS ARE PRESENTED AS EXTRA INFORMATION AND SO AS TO NOT LIMIT YOUR CHOICE OF A CAREER NOT REPRESENTED BY YOUR SEX.

-----YOUR SCORES ON SCALES DEVELOPED ON FEMALE OCCUPATIONAL SAMPLES-----

60	INSTRUMENT ASSEMBLER	59	ARMY OFFICER	58	PHYS. ED. TEACHER	57	X-RAY TECHNICIAN
56	OCCUPATIONAL THERAP.	55	ENGINEER	55	CHEMIST	54	MEDICAL THECHNICIAN
53	PHARMACIST	52	DENTIST	51	DENTAL HYGIENIST	50	PHYS. THERAPIST
49	MATH-SCI. TEACHER	48	DIETITIAN	47	COMPUTER PROGRAMMER	46	MATHEMATICIAN
45	PHYSICIST	44	VETERINARIAN	43	OPTOMETRIST	42	PHYSICIAN
41	COLLEGE PROFESSOR	40	SPEECH PATHOLOGIST	39	PSYCHOLOGIST	38	LANGUAGE INTERPRETER
37	ADVERTISING EXEC.	36	ARTIST	35	ART TEACHER	34	ENTERTAINER
33	INTERIOR DECORATOR	33	LANGUAGE TEACHER	33	LIBRARIAN	32	REPORTER
31	ENGLISH TEACHER	30	NURSE, REGISTERED	29	SOCIAL WORKER	28	DIR. CHRIST. EDUCAT.
27	YWCA STAFF	26	ELEMENTARY TEACHER	25	RECREATION LEADER	24	GUIDANCE COUNSELOR
23	SOC. SCI. TEACHER	22	HOME ECON. TEACHER	21	FLIGHT ATTENDANT	20	LIFE INS. AGENT
10	LAWYER	5	BUYER	0	BANKER	-1	CREDIT MANAGER
-9	DEPT. STORE SALES	-10	BUSINESS ED. TEACHER	-11	EXEC. HOUSEKEEPER	-11	ACCOUNTANT
-12	SECRETARY	-12	DENTAL ASSISTANT	-13	NURSE, LIC. PRAC.	-14	BEAUTICIAN

FIGURE 18

EXAMPLES OF ADMINISTRATIVE SUMMARIES AND NON-OCCUPATIONAL SCALES

C THE FOLLOWING ARE THE ADMINISTRATIVE INDICES AND SPECIAL SCALES. DATA FOR I-VII ARE YOUR LIKE, INDIFFERENT, AND DISLIKE RESPONSE PERCENTAGES (LP-IP-DP) TO THE SEVEN SECTIONS OF THE INVENTORY. TOTAL RESPONSES AND INFREQUENT RESPONSES ARE INTERNAL CHECKS AND THE RESULTS FOR THE SPECIAL SCALES ARE SUMMARIZED BELOW.

	RESPONSE PERCENTAGES		
	LP	IP	DP
I. OCCUPATIONS	80	16	4
II. SCHOOL SUBJECTS	33	33	33
III. ACTIVITIES	99	01	0
IV. AMUSEMENTS	33	33	34
V. TYPES OF PEOPLE	33	33	33
VI. PREFERENCES	20	20	60
VII. CHARACTERISTICS	20	20	60

TOTAL RESPONSES = 325

INFREQUENT RESPONSES = 8

AOR = 55

IE = 35

YOUR SCORE ON THE ACADEMIC ORIENTATION SCALE (AOR) INDICATES A SIMILARITY OF YOUR INTERESTS WITH PEOPLE WHO GO TO COLLEGE AND OBTAIN AN ADVANCED DEGREE BEYOND THEIR BACCALAUREATE DEGREE, SUCH AS DOCTORS, LAWYERS, SCIENTISTS, AND PROFESSORS.

YOUR SCORE ON THE INTROVERSION-EXTRAVERSION SCALE (IE) INDICATES A LIKING FOR WORKING WITH PEOPLE RATHER THAN THINGS. SALES AND SOCIAL SERVICE OCCUPATIONS SCORE ABOUT THE SAME AS YOU DO.

Non-Occupational Scales

Many of the special scales that appeared on the SVIB were not retained on the Strong-Campbell. Older SVIB special scales were eliminated for several reasons: some were based on previous SVIB data that could not be reconstructed successfully, such as the Managerial Orientation (MO) scale; and some have not proven useful, such as the Specialization Level (SL) scale. Campbell (1974) documents the rationale behind the elimination of each of these scales.

Two special scales, that were not eliminated because they have been useful in counseling situations and could be reconstructed for the SCII, were the Academic Orientation Scale (AOR) and the Introversion-Extroversion scale (IE). In both instances, items were used from the original SVIB male and female scales that had been retained in the construction of the SCII; identical scales for males and females were developed and normed on composite samples of males and females.

Academic Orientation (AOR)

The original scale, labeled Academic Achievement, was developed by Campbell and Johansson (1966) for the SVIB to measure interest preferences that were related to academic success in high school and college. The change in scale name to Academic Orientation (AOR) more accurately reflects the purpose of the scale, that of academic inclination or interest; the achievement scale was not intended, nor meant, to be a good predictor of grades, but rather a measure of interest patterns associated with persistence in classroom activities.

Items on the scale that reflect academic achievement deal with scientific and intellectual endeavors, aesthetic interests, and concern for personal competencies. Items that reflect lower academic achievement were fewer in number and centered around clusters of sales and blue-collar types of activities and interests.

Validity data consistently show that when subjects are separated by educational attainment as measured by amount of college experience, higher average scores on AOR correspond to highest educational persistence. The AOR scale for the SCII generally shows that most people with advanced degrees score fairly high on the scale, for example, in the high 50s and greater, while high school drop-outs score very low, less than 30. The interpretive report provides narrative statements for the AOR scale in addition to the actual score which are based on the item content of scales and typical scores for various educational levels.

Introversion-Extroversion Scale (IE)

The original occupational introversion-extroversion scale for the SVIB was developed by contrasting item responses of subjects who were identified as introverted or extroverted. A subject's score on the Social Introversion (Si) scale of the Minnesota Multiphasic Personality Inventory was used as the criterion variable (Johansson, 1970).

The Introversion-Extroversion (IE) scale for the SCII contains those items that survived the merger of the SVIBs. Scores on the IE scale reflect a dimension of preference for working alone versus working with a group of people. High scores are earned by introverts; generally, they prefer to work by themselves with things or ideas. Low scores are earned by extroverts; generally, they prefer to be with a group or speak to a group and enjoy working with others, especially in social service settings. Average scores reflect a balance of interests between the two extremes, neither total acceptance nor rejection of working alone or with others. Interpretive narrative statements provide a description of what various scores mean.

INTERPRETIVE OUTPUT -- CLOSING COMMENTS

The last page of the interpretive report, page 14, contains closing comments as illustrated in Figure 19. The individual is cautioned against overinterpreting the results or against using the scores as hard and fast indicators of an occupational choice. Statements indicate that interest scores are quite stable, and that high scores are related more to job satisfaction and staying in the occupation than are low scores.

Insert Figure 19 About Here

FIGURE 19

CLOSING COMMENTS

YOU HAVE BEEN SCORED ON A BROAD RANGE OF GENERAL INTERESTS AND SPECIFIC OCCUPATIONS. YOU SHOULD NOT BECOME DEAD SET ON ANY ONE PARTICULAR OCCUPATION WHERE YOUR SCORE IS HIGH, AT LEAST NOT AT AN EARLY AGE. IN THE WORLD OF WORK THERE ARE HUNDREDS OF SPECIALITIES AND PROFESSIONS, AND YOU SHOULD USE THIS INFORMATION AS A GUIDE FOR FURTHER EXPLORATION.

THE RESULTS PRESENTED FOR YOU HAVE BEEN BASED ON YEARS OF RESEARCH ON INTEREST PREFERENCES AND ANALYSES OF HUNDREDS OF THOUSANDS OF PEOPLE. STUDIES MADE YEARS LATER OF EMPLOYED PEOPLE WHO COMPLETED EARLIER VERSIONS OF THIS FORM IN THEIR HIGH SCHOOL OR COLLEGE DAYS HAVE SHOWN THAT ABOUT ONE-HALF END UP IN OCCUPATIONS COMPATIBLE WITH THEIR SCORES, AND MOST OF THESE LIKE THEIR WORK. AMONG THOSE WHO END UP IN OCCUPATIONS NOT COMPATIBLE WITH THEIR RESULTS, MANY SAY THEY DO NOT LIKE THEIR WORK, OR THEY ARE DOING THE JOB IN SOME UNUSUAL MANNER.

FURTHER RESEARCH HAS SHOWN THAT SCORES ON INTEREST INVENTORIES ARE VERY STABLE. ADULTS WILL SHOW VERY LITTLE CHANGE IN THEIR SCORES IF TESTED OVER A NUMBER OF YEARS. HIGH SCHOOL STUDENTS WILL SHOW SOME CHANGE IF TESTED AND RETESTED AFTER TWO OR THREE YEARS AND MORE CHANGE AFTER TEN OR FIFTEEN YEARS. GENERALLY, SCORES ON INTEREST INVENTORIES ARE VERY STABLE.

EACH PERSON IS UNIQUE AND NO TEST CAN PREDICT WITH PERFECT ACCURACY ALL THE DIVERSITY AMONG VARIOUS INDIVIDUALS. THUS, THE RESULTS PRESENTED TO YOU SHOULD BE USED AS GUIDELINES IN HELPING YOU TO BETTER UNDERSTAND YOUR CAREER INTERESTS.

FOR ADDITIONAL INFORMATION THE FOLLOWING SOURCES ARE RECOMMENDED--

--CONSULT A PROFESSIONALLY TRAINED COUNSELOR
--THE HANDBOOK FOR THE STRONG VOCATIONAL INTEREST BLANK ($20.00) WRITE STANFORD UNIVERSITY PRESS
--THE MANUAL FOR THE STRONG-CAMPBELL INTEREST INVENTORY ($5.00) WRITE NATIONAL COMPUTER SYSTEMS

061374 DR. CHARLES B. JOHANSSON, NATIONAL COMPUTER SYSTEMS (NCS), 4401 W. 76TH STREET, MINNEAPOLIS, MINNESOTA 55435. ADDITIONAL TESTING MATERIALS MAY BE PURCHASED FROM NCS.

COPYRIGHT 1933 (RENEWED 1961), 1938 (RENEWED 1965), 1945 (RENEWED 1974), 1959, 1964, 1965, 1966, 1968, 1969, 1972, 1974 BY THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY. ALL RIGHTS RESERVED. PRINTED UNDER LICENSE FROM STANFORD UNIVERSITY PRESS, STANFORD, CALIFORNIA 94305.

******END OF REPORT FOR JOHANSSON CHARLES ****** 6/31/74 004 0556

REFERENCES

Campbell, D.P. Handbook for the Strong Vocational Interest Blank. Stanford: Stanford University Press, 1971.

Campbell, D.P. Manual for the Strong-Campbell Interest Inventory. Stanford: Stanford University Press, 1974.

Campbell, D.P., Borgen, F.H., Eastes, S.H., Johansson, C.B., and Peterson, R.A. A set of basic interest scales for the Strong Vocational Interest Blank. Journal of Applied Psychology, 1968, 52, 6, part 2.

Campbell, D.P. and Holland, J.L. A merger in vocational interest research: applying Holland's theory to Strong's data. Journal of Vocational Behavior, 1972, 2, 353-376.

Campbell, D.P. and Johansson, C.B. Academic interests, scholastic achievement, and eventual occupations. Journal of Counseling Psychology, 1966, 13, 4, 416-424.

Hansen, J.C. and Johansson, C.B. The application of Holland's theory to the Strong Vocational Interest Blank for women. Journal of Vocational Behavior. 1972, 2, 479-493.

Holland, J.L. The psychology of vocational choice. Waltham: Blaisdell Publishing, 1966.

Holland, J.L. Making vocational choices: a theory of careers. Englewood Cliffs, Prentice-Hall, Inc., 1973.

Johansson, C.B. Strong Vocational Interest Blank introversion-extroversion and occupational membership. Journal of Counseling Psychology, 1970, 17, 5, 451-455.

Johansson, C.B. Cognitive interest styles of students. Journal of Measurement and Evaluation in Guidance, 1971, 4, 3, 176-184.

Roberts, C.A. and Johansson, C.B. The inheritance of cognitive interest styles among twins. Journal of Vocational Behavior, 1974, 4, 237-243.

United States Department of Labor. Dictionary of occupational titles. Occupational Classification and Industrial Index, 1965, II, Third Edition.

United States Department of Labor. Occupational outlook handbook (1974-1975). Bureau of Labor Statistics, Bulletin 1785.

APPENDIX 1

EXAMPLE OF INTERPRETIVE REPORT

999 0025

STRONG-CAMPBELL INTEREST INVENTORY

TEST RESULTS FOR NO NAME 10/16/74 MALE

THESE TEST RESULTS ARE FROM DATA YOU PROVIDED WHEN YOU ANSWERED THE STRONG-CAMPBELL INTEREST INVENTORY. THE RESULTS ARE BASED ON YOUR LIKE AND DISLIKE ANSWERS TO THE INVENTORY AND THEY ARE IMPORTANT IN HELPING YOU TO BETTER UNDERSTAND HOW YOUR PREFERENCES FIT INTO THE WORLD OF WORK. THEY POINT OUT WORK AREAS WHERE YOUR INTERESTS MATCH THOSE OF EMPLOYED PEOPLE. THEY ALSO CAN BE USEFUL IN INDICATING POTENTIAL PROBLEMS WHERE YOUR INTERESTS DIFFER SUBSTANTIALLY FROM PEOPLE IN OCCUPATIONS THAT YOU MIGHT BE CONSIDERING.

IMPORTANTLY, THESE RESULTS ARE MEASURES OF YOUR INTERESTS AND NOT OF YOUR ABILITIES. FOR EXAMPLE, THE SCORES MAY INDICATE THAT YOU WOULD LIKE THE DAILY ROUTINE OF ARTISTS AND THAT YOU LIKE ARTISTIC ACTIVITIES, BUT THEY WILL NOT TELL YOU IF YOU HAVE THE TALENT TO BE AN ARTIST.

THE RESULTS CAN GIVE YOU SOME USEFUL INFORMATION ABOUT YOURSELF, BUT DO NOT EXPECT MIRACLES. ALTHOUGH YOU MAY FEEL THAT THE SCORES ACTUALLY TELL YOU NOTHING MORE THAN YOU ALREADY KNEW, THEY WILL TELL YOU HOW HIGH OR LOW YOUR INTERESTS ARE IN RELATION TO THE AVERAGE SCORE OF OTHERS.

* * * * * * * * * * * * * * * * * * * *

THREE MAIN SETS OF SCORES ARE PRESENTED BELOW. THE FIRST SET OF SCORES ARE YOUR RESULTS ON SIX GENERAL OCCUPATIONAL THEMES--THEY GIVE YOU A GENERAL OVERALL VIEW OF YOUR INTERESTS. THE SECOND SET OF SCORES ARE YOUR RESULTS ON THE BASIC INTEREST SCALES--THESE SCORES TELL YOU ABOUT THE STRENGTH OF YOUR INTERESTS IN MORE SPECIFIC AREAS SUCH AS SALES, MUSIC, AND MECHANICAL. THE THIRD SET OF SCORES ARE THE OCCUPATIONAL SCALES--THEY TELL YOU HOW SIMILAR OR DISSIMILAR YOUR INTERESTS ARE TO EMPLOYED PEOPLE IN VARIOUS OCCUPATIONS SUCH AS ARTISTS OR MATHEMATICIANS.

* * * * * * * * * * * * * * * * * * * *

```
                 ***********************************************
                 *****                                     *****
                 *****     GENERAL OCCUPATIONAL THEMES     *****
                 *****                                     *****
                 ***********************************************
```

PSYCHOLOGICAL RESEARCH HAS SHOWN THAT INTERESTS CAN BE GROUPED INTO SIX CATEGORIES AND THAT THESE GROUPINGS CAN BE DESCRIBED BY A GENERAL THEME. IMMEDIATELY BELOW ARE YOUR SCORES ON THE SIX GENERAL THEMES AND A GRAPH TO VIEW YOUR RESULTS BETTER. AN AVERAGE PERSON IN THE GENERAL POPULATION SCORES BETWEEN 43 AND 57 ON THE SCALES. SCORES BELOW 43 GENERALLY ARE CONSIDERED AS LOW AND SCORES ABOVE 57 GENERALLY ARE CONSIDERED AS HIGH.

```
           YOUR      3         4         5         6         7
 SCALE     SCORE ....0.........0.........          0.........0....

 I-THEME    69   *********************************************      (INVESTIGATIVE)
 A-THEME    61   ************************************               (ARTISTIC)
 R-THEME    56   *******************************                    (REALISTIC)
 C-THEME    48   ***********************                            (CONVENTIONAL)
 S-THEME    42   *****************                                  (SOCIAL)
 E-THEME    38   *************                                      (ENTERPRISING)
```

ALTHOUGH DESCRIPTIONS ARE GIVEN BELOW FOR EACH THEME, THEY DESCRIBE A GROUP OF PEOPLE AND ALL THE CHARACTERISTICS MAY NOT FIT EXACTLY ANY ONE PERSON. FEW PEOPLE SCORE HIGH ON JUST ONE THEME AND LOW ON ALL OTHERS. MOST SCORE HIGH ON TWO OR EVEN THREE THEMES AND THUS THEY SHARE SOME OF THE CHARACTERISTICS WITH MORE THAN ONE TYPE. SOME PEOPLE SCORE LOW ON ALL THEMES WHICH WOULD INDICATE THAT DEFINITE PATTERNS OF INTERESTS, AS MEASURED BY THESE THEMES, HAVE NOT BEEN ESTABLISHED YET--THIS IS PARTICULARLY TRUE FOR YOUNG PEOPLE. GENERALLY, THE HIGHER YOUR SCORE, THE MORE CHARACTERISTICS YOU SHARE WITH THAT PARTICULAR THEME. MEN AND WOMEN SCORE SOMEWHAT DIFFERENTLY ON THE SAME THEME AND THIS HAS BEEN TAKEN INTO ACCOUNT BY COMPARING YOUR SCORES WITH THE AVERAGES FOR YOUR SEX.

```
I      I-THEME---THIS THEME TENDS TO CENTER AROUND SCIENCE AND SCIENTIFIC ACTIVITIES.  HIGH          I
N  SCORERS ON THIS THEME WOULD PREFER TO WORK ALONE RATHER THAN IN A LARGE OFFICE WITH A GROUP OF    N
V  PEOPLE.  THEY PREFER TO THINK THROUGH PROBLEMS AND WORK WITH IDEAS, WORDS, AND SYMBOLS IN SOL-    V
E  VING THEM.  SUCH PEOPLE ENJOY LOOSELY-DEFINED PROBLEMS AND DO NOT LIKE HIGHLY-STRUCTURED SITUA-   E
S  TIONS WITH MANY RULES.  THEY FREQUENTLY TEND TO BE ORIGINAL AND CREATIVE, ESPECIALLY IN          S
T  SCIENTIFIC AREAS.  THEY PREFER OCCUPATIONS SUCH AS DESIGN ENGINEER, MATHEMATICIAN, BIOLOGIST,     T
I  PSYCHOLOGIST, RESEARCH LABORATORY WORKER, PHYSICIST, PHYSICIAN, TECHNICAL WRITER, OR METEOROL-    I
G  OGIST.  THE WORD INVESTIGATIVE IS USED TO SUMMARIZE THIS PATTERN, HENCE I-THEME.                 G
.      YOUR SCORE OF 69 ON THE I-THEME SCALE INDICATES THAT VERY LIKELY MOST OF THE ABOVE DESCRIP-   .
   TIONS ARE TRUE FOR YOU AND YOU WOULD ENJOY ACTIVITIES THAT INVOLVE SCIENTIFIC PROBLEMS.
   IN COMPARISON WITH OTHER SCALES, YOUR SCORE IS A VERY HIGH SCORE.
```

ARTISTIC

A-THEME---THE A-THEME IS ARTISTICALLY ORIENTED AND LIKES TO WORK IN ARTISTIC SETTINGS WHERE
THERE ARE MANY OPPORTUNITIES FOR SELF-EXPRESSION. SUCH PEOPLE HAVE LITTLE INTEREST IN SITUA-
TIONS THAT ARE VERY STRUCTURED OR REQUIRE PHYSICAL STRENGTH. INSTEAD THEY PREFER PROBLEMS OR
TASKS THAT CAN BE DEALT WITH THROUGH SELF-EXPRESSION IN ARTISTIC MEDIA. THEY PREFER TO WORK
ALONE AND HAVE A STRONG NEED FOR INDIVIDUAL EXPRESSION. THEY SCORE HIGHER ON MEASURES OF ORIG-
INALITY THAN ANY OF THE OTHER TYPES. THEY DESCRIBE THEMSELVES AS INDEPENDENT, UNCONVENTIONAL,
ORIGINAL, EXPRESSIVE, AND INTUITIVE. VOCATIONAL CHOICES INCLUDE ARTIST, AUTHOR, CARTOONIST, COMPOS-
ER, SINGER, DRAMATIC COACH, POET, ACTOR OR ACTRESS, AND SYMPHONY CONDUCTOR. THIS IS THE ARTIS-
TIC THEME, OR A-THEME.

YOUR SCORE OF 61 ON THE A-THEME SCALE INDICATES THAT MANY OF THE ABOVE DESCRIPTIONS ARE
MOST LIKELY TRUE FOR YOU AND YOU MIGHT FIND ARTS AND CRAFTS TYPES OF ACTIVITIES REWARDING.
IN COMPARISON WITH OTHER MALES, YOUR SCORE IS A HIGH SCORE.

REALISTIC

R-THEME---DESCRIPTIVE EXAMPLES ARE RUGGED, PRACTICAL, AND PHYSICALLY STRONG. SUCH PEOPLE
USUALLY HAVE TROUBLE EXPRESSING THEMSELVES IN WORDS OR COMMUNICATING THEIR FEELINGS TO OTHERS.
THEY LIKE TO WORK OUTDOORS, AND THEY LIKE TO WORK WITH TOOLS, ESPECIALLY LARGE MACHINES. THEY
PREFER TO DEAL WITH OBJECTS RATHER THAN WITH IDEAS OR WITH PEOPLE. THEY ENJOY CREATING THINGS
WITH THEIR HANDS AND PREFER OCCUPATIONS SUCH AS MECHANIC, CONSTRUCTION WORK, FISH AND WILDLIFE
MANAGEMENT, LABORATORY TECHNICIAN, SOME ENGINEERING SPECIALTIES, SOME MILITARY JOBS, AGRICULTURE
OR THE SKILLED TRADES. ALTHOUGH NO SINGLE WORD CAN BEST CAPTURE THE BROAD MEANING OF THE ENTIRE
THEME, THE WORD REALISTIC HAS BEEN USED, THUS R-THEME.

YOUR SCORE OF 56 ON THE R-THEME SCALE INDICATES SOME OF THE ABOVE PHRASES ARE PROBABLY
TRUE FOR YOU, WHILE SOME OF THE OTHER PHRASES PROBABLY DO NOT DESCRIBE YOU AS WELL.
IN COMPARISON WITH OTHER MALES, YOUR SCORE IS AN AVERAGE SCORE.

CONVENTIONAL

C-THEME---HIGH SCORERS ON THIS TYPE PREFER JOBS WHERE THEY KNOW EXACTLY WHAT THEY ARE SUP-
POSED TO DO, USING VERBAL AND NUMERICAL SKILLS THAT CHARACTERIZE OFFICE WORK. THEY FIT WELL
INTO LARGE ORGANIZATIONS, BUT DO NOT SEEK LEADERSHIP. THEY RESPOND TO POWER AND ARE COMFORTABLE
WORKING IN A WELL-ESTABLISHED CHAIN OF COMMAND. THEY DISLIKE LOOSELY-DEFINED JOB SITUATIONS,
PREFERRING TO KNOW EXACTLY WHAT IS EXPECTED OF THEM. SUCH PEOPLE DESCRIBE THEMSELVES AS CONVEN-
TIONAL, STABLE, WELL-CONTROLLED, AND DEPENDABLE. THEY HAVE LITTLE INTEREST IN PROBLEMS THAT RE-
QUIRE PHYSICAL SKILLS OR INTENSE RELATIONSHIPS WITH OTHERS. THEY ARE MOST EFFECTIVE AT WELL-
DEFINED TASKS, AND THEY VALUE MATERIAL POSSESSIONS AND STATUS. VOCATIONAL PREFERENCES ARE MOST-
LY WITHIN THE BUSINESS WORLD AND INCLUDE BANK EXAMINER, BANK TELLER, BOOKKEEPER, SOME ACCOUNTING
JOBS, FINANCIAL ANALYST, COMPUTER OPERATOR, INVENTORY CONTROLLER, TAX EXPERT, STATISTICIAN, AND
TRAFFIC MANAGER. ALTHOUGH ONE WORD CANNOT ADEQUATELY REPRESENT THE ENTIRE THEME, THE WORD CON-
VENTIONAL SUMMARIZES THIS PATTERN, HENCE C-THEME.

YOUR SCORE OF 48 ON C-THEME INDICATES THAT SOME OF THE ABOVE PHRASES ARE TRUE FOR YOU.
IN COMPARISON WITH OTHER MALES, YOUR SCORE IS AN AVERAGE SCORE.

SOCI

S-THEME---PEOPLE WHO SCORE HIGH ARE SOCIABLE, RESPONSIBLE, AND CONCERNED WITH THE WELFARE
OF OTHERS AND HUMAN PROBLEMS. THEY USUALLY EXPRESS THEMSELVES WELL AND GET ALONG WITH OTHERS.
THEY LIKE TO SEEK ATTENTION AND SITUATIONS THAT WILL ALLOW THEM TO BE AT OR NEAR THE CENTER OF
THE GROUP. THEY PREFER TO SOLVE PROBLEMS BY DISCUSSIONS WITH OTHERS, OR BY ARRANGING OR
REARRANGING RELATIONSHIPS BETWEEN OTHERS. THEY HAVE LITTLE INTEREST IN SITUATIONS REQUIRING
PHYSICAL ACTIVITY OR WORKING WITH MACHINERY. SUCH PEOPLE DESCRIBE THEMSELVES AS CHEERFUL, POP-

A
L

ULAR, ACHIEVING, AND GOOD LEADERS. THEY PREFER OCCUPATIONS SUCH AS SCHOOL SUPERINTENDENT, CLINICAL PSYCHOLOGIST, HIGH SCHOOL TEACHER, MARRIAGE COUNSELOR, PLAYGROUND DIRECTOR, SPEECH THERAPIST, OR VOCATIONAL COUNSELOR. THIS IS THE SOCIAL THEME, OR S-THEME.

YOUR SCORE OF 42 ON THE S-THEME INDICATES THAT PROBABLY VERY FEW OF THE ABOVE DESCRIPTIONS ARE TRUE FOR YOU AND YOU MAY NOT ENJOY WORKING WITH PEOPLE OR HELPING THEM WITH THEIR PROBLEMS. IN COMPARISON WITH OTHER MALES, YOUR SCORE IS MODERATELY LOWER THAN AVERAGE.

A
L

E
N
T
E
R
P
R
I
S
I
N
G

E-THEME---THE HIGH SCORERS ARE GOOD WITH WORDS WHICH THEY PUT TO EFFECTIVE USE IN SELLING, DOMINATING, AND LEADING. FREQUENTLY THEY ARE IN SALES WORK, THEY ARE GOOD AT THINKING UP NEW WAYS OF DOING THINGS AND AT LEADING AND CONVINCING PEOPLE. THEY SEE THEMSELVES AS FULL OF ENERGY, ENTHUSIASTIC, LIKING ADVENTURE, SELF-CONFIDENT, AND DOMINANT. THEY PREFER SOCIAL SITUATIONS WHERE THEY CAN ASSUME LEADERSHIP AND DIRECT OTHERS. THEY ENJOY PERSUADING OTHERS TO THEIR POINT OF VIEW. THEY ARE IMPATIENT WITH PRECISE WORK OR WORK INVOLVING LONG PERIODS OF INTELLECTUAL EFFORT OR CONCENTRATION FOR LONG PERIODS OF TIME. THEY LIKE POWER, STATUS, AND MATERIAL WEALTH, AND ENJOY WORKING IN EXPENSIVE SETTINGS. VOCATIONAL PREFERENCES INCLUDE BUSINESS EXECUTIVE, BUYER, HOTEL MANAGER, INDUSTRIAL RELATIONS CONSULTANT, POLITICAL CAMPAIGNER, REALTOR, MANY KINDS OF SALES WORK, SPORTS PROMOTER, AND TELEVISION PRODUCER. THE WORD ENTERPRISING SUMMARIZES THIS PATTERN OF INTERESTS, THUS E-THEME.

YOUR SCORE OF 38 ON THE E-THEME SCALE INDICATES THAT PROBABLY VERY FEW OF THE ABOVE DESCRIPTIONS ARE TRUE FOR YOU. POSSIBLY, BUSINESS ACTIVITIES OR SELLING WOULD NOT BE ENJOYABLE. IN COMPARISON WITH OTHER MALES, YOUR SCORE IS A LOW SCORE.

E
N
T
E
R
P
R
I
S
I
N
G

* * * * * * * * * * * * * * * * * * * *

THE ABOVE HAS BEEN AN ANALYSIS OF YOUR INTERESTS ON SIX GENERAL THEME SCALES R-I-A-S-E-C. THESE SCALES ARE BASED ON THE PSYCHOLOGICAL WORKS OF JOHN HOLLAND AND FURTHER INFORMATION ABOUT THE THEMES CAN BE FOUND IN HIS EXCELLENT LITTLE BOOK--MAKING VOCATIONAL CHOICES, A THEORY OF CAREERS, BY PRENTICE-HALL, ENGLEWOOD CLIFFS, NEW JERSEY, FOR $4.95(PAPERBACK).

THE PRECEDING R-I-A-S-E-C SCORES PROVIDED YOU WITH A BROAD, GENERAL, OVER-VIEW OF YOUR INTERESTS. THE NEXT SECTION PRESENTS YOUR RESULTS ON THE MORE SPECIFIC AND LESS GENERAL BASIC INTEREST SCALES. THESE SCALES GIVE YOU MORE SPECIFIC INFORMATION ON YOUR INTERESTS IN PURE TYPES OF ACTIVITIES, SUCH AS ART OR TEACHING OR WORKING WITH NUMBERS.

* * * * * * * * * * * * * * * * * * * *

```
                                                                                                          P. 1
                                     *******************************************
                                     *****      BASIC INTEREST SCALES      *****
                                     *******************************************
     BELOW ARE YOUR RESULTS ON 23 BASIC INTEREST SCALES.  THEY SHOW THE STRENGTH OF YOUR INTER-
ESTS IN A VARIETY OF PURE TYPES OF AREAS.  THESE SCORES WILL SHOW AREAS WHERE YOU WILL FIND SAT-
ISFACTION OR DISSATISFACTION.  THEY CAN BE IMPORTANT IN MAKING AN OCCUPATIONAL CHOICE OR IN DE-
CIDING ON LEISURE-TIME ACTIVITIES.  THE AVERAGE ADULT SCORES BETWEEN 43 AND 57 ON EACH SCALE.
HIGH SCORES SHOW AREAS WHERE YOU WILL FIND ENJOYABLE TYPES OF ACTIVITIES.  LOW SCORES INDICATE
AREAS THAT WILL BE LESS REWARDING TO YOU.  ON SOME SCALES MALES AND FEMALES RESPOND SOMEWHAT
DIFFERENTLY AND THE RANGE OF DOTS INDICATES THE AVERAGE RANGE OF SCORES FOR YOUR SEX.  ALSO, THE
LOWEST POSSIBLE SCORE (MIN) AND HIGHEST POSSIBLE SCORE (MAX) ARE LISTED.
                          YOUR       3         4         5         6         7
BASIC INTEREST SCALES     SCORE  ....'.........'.........'.........'.........'....   MIN-MAX

****VERY HIGH (66+)****
SCIENCE                   69                        ..............          *        28-69
MATHEMATICS               67                        ..............        *          31-67

****HIGH (65-58)****
MEDICAL SCIENCE           63                       .............      *              29-68
MECHANICAL                59                           ...........*..                30-72
AGRICULTURE               58                        .............*                   30-68
NATURE                    58                    ...............  *                   17-67

****AVERAGE (57-43)****
ART                       56                 ................  *                     26-68
WRITING                   55                   ...............*                      26-66
MUSIC/DRAMA               54                  ...............*                       26-73
ADVENTURE                 53                          ......*.......                 30-71
TEACHING                  52                    ...........*...                      23-67
MEDICAL SERVICE           50                     ........*...                        33-76
PUBLIC SPEAKING           49                        ....*..........                  31-71
LAW/POLITICS              46                         *.............                  30-69
DOMESTIC ARTS             44               ........*.....                            27-75
OFFICE PRACTICES          44                    ...*.........                        36-79
SOCIAL SERVICE            43                   ...*............                      24-69

****LOW (42-35)****
MILITARY ACTIVITIES       41                    *   ................                 41-76
BUSINESS MANAGEMENT       41                    *   ...............                  29-72
ATHLETICS                 39                  *      ...............                 27-70
SALES                     37                *       ................                 37-81

****VERY LOW (34-)****
RELIGIOUS ACTIVITIES      35              *      ..............                      31-68
MERCHANDISING             34             *        ...............                    31-72
                                     3         4         5         6         7
                                 ....'.........'.........'.........'.........'....
```

* * *

YOUR HIGHEST SCORES ON THE BASIC INTEREST SCALES ARE LISTED BELOW WITH EXAMPLES OF OCCUPATIONS THAT SCORE HIGH AND LOW. PEOPLE IN OCCUPATIONS THAT ARE EXAMPLES OF HIGH SCORES LIKE THE ACTIVITIES THAT ARE MEASURED BY THAT SCALE. PEOPLE IN OCCUPATIONS THAT ARE EXAMPLES OF THE LOW SCORES HAVE STRONG DISLIKES FOR THOSE SAME ACTIVITIES. THE HIGHER YOUR SCORE, THE MORE SIMILAR YOUR INTERESTS ARE TO PEOPLE IN THE HIGH SCORING OCCUPATIONS AND THE MORE DISSIMILAR YOUR INTERESTS ARE TO PEOPLE IN OCCUPATIONS THAT SCORE LOW.

ALSO, PAGE REFERENCES ARE PRESENTED FOR THE OCCUPATIONAL OUTLOOK HANDBOOK (OOH), BULLETIN NO. 1785 THAT CAN BE USED FOR FURTHER REFERENCE. THE OOH IS A GOOD SOURCE FOR RECENT TRENDS IN EMPLOYMENT OPPORTUNITIES AND EDUCATIONAL REQUIREMENTS FOR VARIOUS OCCUPATIONS THAT ARE RELATED TO YOUR HIGH AND LOW SCORES. THE OOH CAN BE FOUND IN YOUR LOCAL LIBRARY OR ORDERED DIRECTLY FROM THE U.S. GOVERNMENT PRINTING OFFICE, WASHINGTON, D.C., 20402 ($6.85).

YOUR ---HIGHEST SCORES---	----EXAMPLES OF HIGH SCORES-----	-----EXAMPLES OF LOW SCORES-----
69 SCIENCE OOH 329-401	ASTRONOMER, COLLEGE PROFESSOR, PHYSICIAN, BIOLOGIST, CHEMIST	BUSINESS EXECUTIVE, BEAUTICIAN, ARTIST, WRITER, SALES PERSONNEL
67 MATHEMATICS OOH 371-376	MATH-SCIENCE TEACHER, ASTRONAUT, MATHEMATICIAN, ENGINEER, CHEMIST	INTERIOR DECORATOR, REPORTER, FASHION MODEL, ARTIST
63 MEDICAL SCIENCE OOH 473-484	MEDICAL TECHNOLOGIST, PHYSICIAN, PHYSICAL THERAPIST, VETERINARIAN	SEWING MACHINE OPERATOR, FARMER WRITER, ARTIST

YOUR HIGHEST SCORES ON THE BASIC INTEREST SCALES INDICATE YOUR STRONGEST PREFERENCE FOR THE PURE TYPES OF INTERESTS AND ACTIVITIES MEASURED BY THESE SCALES. IN GENERAL, THESE AREAS PROBABLY WILL BE SOURCES OF SATISFACTION IN YOUR LIFE. ALTHOUGH THEY MAY OR MAY NOT BE PART OF YOUR OCCUPATIONAL CHOICE, YOU PROBABLY WILL FIND YOUR OCCUPATIONAL CHOICE MORE REWARDING IF THE INTERESTS ARE INCLUDED. YOUR HIGH SCORES ALSO WILL POINT OUT AREAS OF SATISFACTION FOR LEISURE TIME ACTIVITIES.

TO GIVE YOU MORE INFORMATION ABOUT YOUR HIGH SCORES, THE FOLLOWING GENERAL STATEMENTS WILL SUMMARIZE THE IMPORTANT PARTS OF THESE SCORES. GENERALLY, THE HIGHER YOUR SCORE THE MORE APPLICABLE THE PHRASES ARE FOR YOU. YOU HAVE INDICATED THAT YOU LIKE--

--OCCUPATIONS RELATED TO SCIENTIFIC INVESTIGATION AND RESEARCH WORK.
--STUDIES RELATED TO COMPUTATIONAL AND NUMERICAL PROBLEMS.
--OCCUPATIONS RELATED TO PREVENTING ILLNESSES AND MAINTAINING HEALTH.

* * *

THE FOLLOWING ARE YOUR LOWEST SCORES ON THE BASIC INTEREST SCALES. THEY INDICATE AREAS WHERE YOUR INTEREST PREFERENCES ARE LEAST STRONG FOR THE ACTIVITIES AND AREAS MEASURED BY THESE SCALES. YOU PROBABLY WILL HAVE LESS SATISFACTION WITH A CAREER RELATED DIRECTLY TO YOUR LOWEST

SCORES THAN A CAREER RELATED TO YOUR HIGHEST SCORES. YOUR LOWEST SCORES ARE MOST SIMILAR TO THE LOW SCORING OCCUPATIONS AND LEAST SIMILAR TO THE OCCUPATIONS THAT SCORE HIGH.

YOUR ----LOWEST SCORES---	-----EXAMPLES OF LOW SCORES-----	----EXAMPLES OF HIGH SCORES-----
34 MERCHANDISING OOH 235-237	PHOTOGRAPHER, SCIENTIST, ARTIST, WRITER	BUSINESS EDUCATION TEACHER, BUYER, DEPARTMENT STORE MANAGER
35 RELIGIOUS ACTVS. OOH 553-558	PSYCHOLOGIST, PHYSICIAN, ARTIST, CHEMIST	DIRECTOR, CHRISTIAN EDUCATION, PRIEST, MINISTER, NUN
37 SALES OOH 213-243	ANTHROPOLOGIST, SCIENTIST, ARTIST, WRITER	SALES MANAGER, SALES PERSONNEL, BUYER, BUSINESS EDUCATION TCHR.

YOUR LOWEST SCORES ON THE BASIC INTEREST SCALE INDICATE YOUR WEAKEST PREFERENCES FOR THE PURE TYPES OF INTERESTS AND ACTIVITIES MEASURED BY THESE SCALES. IN GENERAL, YOU PROBABLY WILL FIND OTHER AREAS MORE SATISFYING THAN THOSE INDICATED BY YOUR LOW SCORES.

TO GIVE YOU MORE INFORMATION ABOUT THESE SCORES, THE FOLLOWING GENERAL STATEMENTS WILL SUMMARIZE THE IMPORTANT ASPECTS. GENERALLY, THE LOWER YOUR SCORE THE MORE APPLICABLE THE STATEMENT IS FOR YOU. YOUR ANSWERS TO THE INVENTORY HAVE INDICATED THAT YOU DO NOT LIKE--

--DEALING WITH THE SALES MARKET THROUGH ADVERTISING, BUYING, AND PROMOTING PRODUCTS.
--BEING INVOLVED IN CHURCH-RELATED ACTIVITIES.
--OCCUPATIONS THAT INVOLVE MEETING PEOPLE AND SELLING THINGS TO THEM.

****

THE PRECEDING HAS BEEN AN ANALYSIS OF YOUR INTERESTS ON 23 BASIC INTEREST SCALES. THESE SCALES ALSO ARE RELATED TO THE PREVIOUS 6 GENERAL OCCUPATIONAL THEME SCALES. SUMMARIZED BELOW ARE YOUR SCORES ON EACH THEME SCALE AND THE RELATED BASIC INTEREST SCALES. REMEMBER, IN A GENERAL POPULATION, THE AVERAGE ADULT HAS A SCORE OF 50 ON EACH SCALE. CONSISTENTLY HIGH OR LOW SCORES IN AN AREA SHOULD BE IMPORTANT INDICATORS FOR YOU IN UNDERSTANDING YOUR INTERESTS.

I. 69 I-THEME --- (INVESTIGATIVE)	II. 6[illegible] A-THEME --- (ARTISTIC)	III. 56 R-THEME --- (REALISTIC)
69 SCIENCE	56 ART	59 MECHANICAL
67 MATHEMATICS	55 WRITING	58 AGRICULTURE
63 MEDICAL SCIENCE	54 MUSIC/DRAMA	58 NATURE
5[illegible] MEDICAL SERVICE		53 ADVENTURE
		41 MILITARY ACTIVITIES

IV. 48 C-THEME ---
(CONVENTIONAL)

44 OFFICE PRACTICES

V. 42 S-THEME ---
(SOCIAL)

52 TEACHING
44 DOMESTIC ARTS
43 SOCIAL SERVICE
39 ATHLETICS
35 RELIGIOUS ACTIVITIES

VI. 38 E-THEME ---
(ENTERPRISING)

49 PUBLIC SPEAKING
46 LAW/POLITICS
41 BUSINESS MANAGEMENT
37 SALES
34 MERCHANDISING

****

THE PRECEDING TWO SECTIONS HAVE ANALYZED YOUR RESULTS IN TERMS OF GENERAL OCCUPATIONAL THEME SCALES AND THE MORE SPECIFIC BASIC INTERESTS SCALES. THE FOLLOWING PAGE STARTS THE NEXT SECTION OF ANALYSIS, THE OCCUPATIONAL SCALES. THESE SCALES ARE MORE SPECIFIC STILL. THEY COMPARE YOUR ANSWERS WITH EMPLOYED PEOPLE IN A VARIETY OF OCCUPATIONS.

SINCE MALES AND FEMALES RESPOND DIFFERENTLY TO SOME ITEMS ON THIS INVENTORY, THE RESULTS PRESENTED BELOW ARE FOR THOSE SCALES THAT WERE BASED ON WORKERS WHO WERE OF THE SAME SEX AS YOU. TOWARDS THE END OF THIS REPORT, YOUR RESULTS ARE REPORTED FOR SCALES DEVELOPED ON THE OPPOSITE SEX. HOWEVER, AN INTENSIVE ANALYSIS IS NOT PROVIDED SINCE RESEARCH HAS SHOWN THAT SAME-SEX SCALES ARE MORE VALID FOR AN INDIVIDUAL THAN OTHER-SEX SCALES.

****

OCCUPATIONAL SCALES

THE NEXT GROUP OF RESULTS ARE OCCUPATIONAL SCALES. THEY INDICATE THE DEGREE OF SIMILARITY BETWEEN YOUR INTERESTS AND THOSE OF EMPLOYED PEOPLE WHO ARE OF THE SAME SEX AS YOU. THE AVERAGE WORKER IN AN OCCUPATION SCORES ABOUT 50 ON THAT SCALE AND ABOUT 2/3 OF THESE WORKERS HAVE SCORES BETWEEN 40 AND 60. MOST PEOPLE NOT IN THE OCCUPATION SCORE ABOUT 30 ON THE SCALE.

RESEARCH HAS INDICATED THAT PEOPLE WHO ENTER AN OCCUPATION WHERE THEY HAVE HIGH SCORES TEND TO REMAIN IN THAT OCCUPATION AND ARE MORE SATISFIED THAN IF THEY ENTER AN OCCUPATION WHERE THEY HAVE LOW SCORES. FURTHERMORE, THESE SCORES MAY INDICATE INTEREST IN AN OCCUPATION THAT YOU MAY NOT HAVE CONSIDERED BEFORE AND CAN HELP YOU THINK ABOUT OTHER VARIOUS CAREER POSSIBILITIES.

VERY SIMILAR 55+		SIMILAR 54-45		AVERAGE 44-26		DISSIMILAR 25-16		VERY DISSIMILAR 15-	
[illegible]	PSYCHOLOGIST	54	ARCHITECT	44	COLLEGE PROF.	24	PRIEST	14	COMPUTER SALES
62	COMPUTER PROGR.	51	ENGINEER	44	ELEM. TEACHER	23	DIETITIAN	13	FARMER
61	SPEECH PATHOL.	50	PHOTOGRAPHER	42	MERCH MARINE OFF	23	SKILLED CRAFTS	13	POLICE OFFICER
59	SOCIAL SCIENTIST	49	CARTOGRAPHER	41	OPTOMETRIST	23	MINISTER	13	GUIDANCE COUNS.
57	PHYSICIAN	49	MEDICAL TECH.	38	LIBRARIAN	20	SCHOOL SUPERDT.	13	CREDIT MANAGER
57	MATH-SCI. TEACH.	48	ARTIST	37	AIR FORCE OFF.	19	CHAM. OF COMM.	13	BUSINESS ED TCHR
56	BIOLOGIST	46	VETERINARIAN	37	INVEST. FUND MGR	19	PURCHASING AGENT	11	RECREATION LEAD.
55	PHYS. SCIENTIST	46	DENTIST	36	ENGLISH TEACHER	18	SOCIAL WORKER	11	DEPT. STORE MGR.
55	MATHEMATICIAN	45	NAVY OFFICER	34	FORESTER	17	SOC. SCI TEACHER	8	REALTOR
55	MUSICIAN			34	ADVERTISING EXEC			8	ACCOUNTANT
				34	REPORTER			7	BANKER
				34	PHYS. THERAPIST			6	SALES MANAGER
				32	REGISTERED NURSE			5	VOC AGRIC. TCHR.
				3[illegible]	CHIROPRACTOR			3	HWY. PATROL OFF.
				29	PUBLIC ADMINIS.			2	LIFE INS. AGENT
				28	ARMY OFFICER			0	FUNERAL DIRECTOR
				28	INT. DECORATOR			-2	BUYER
				27	LIC. PRAC. NURSE			-6	AGRIBUSINESS MGR
				27	PERSONNEL DIR.				
				27	PHARMACIST				
				26	LAWYER				

****

BELOW ARE YOUR HIGHEST AND LOWEST OCCUPATIONAL SCORES. YOU SHOULD PAY PARTICULAR ATTENTION TO THOSE OCCUPATIONS WHERE YOUR SCORES INDICATED THAT YOU HAD THE MOST SIMILAR INTERESTS AND TO THOSE OCCUPATIONS WHERE YOUR INTERESTS WERE MOST DISSIMILAR. YOU WILL HAVE THE BEST CHANCE OF FINDING SATISFACTION IF YOU DECIDE ON AN OCCUPATION WHERE YOUR INTERESTS ARE SIMILAR WITH YOUR CO-WORKERS AND LESS CHANCE IF YOUR INTERESTS ARE DISSIMILAR. IF YOU RECORDED THE SAME LIKES AND DISLIKES AS THE WORKERS, YOUR SCORE WILL BE HIGH FOR THAT OCCUPATION. IF YOUR LIKE AND DISLIKE RESPONSES WERE DIFFERENT FROM THOSE OF PEOPLE IN THE OCCUPATION, YOUR SCORE WILL BE LOW AND YOU WOULD NOT LIKELY BE HAPPY IN THAT KIND OF WORK.

AS BEFORE, REFERENCES ARE GIVEN FOR THE OCCUPATIONAL OUTLOOK HANDBOOK (OOH) WHICH PRESENTS ADDITIONAL INFORMATION ON EMPLOYMENT OPPORTUNITIES AND RELEVANT WORK SITUATIONS. ALSO, REFERENCES ARE LISTED FOR THE DICTIONARY OF OCCUPATIONAL TITLES (DOT) VOLUME I, THIRD EDITION. THE DOT GIVES DETAILED JOB DESCRIPTIONS OF THE DUTIES AND FUNCTIONS OF EACH OCCUPATION AND CAN BE FOUND IN YOUR LOCAL LIBRARY OR PURCHASED FROM THE U.S. GOVERNMENT PRINTING OFFICE ($7.75).

YOUR HIGHEST SCORES APPEARED ON THE FOLLOWING SCALES AND INDICATE THE GREATEST DEGREE OF SIMILARITY BETWEEN YOUR ANSWERS AND THOSE OF MALES IN THESE OCCUPATIONS--

	64 PSYCHOLOGIST	62 COMPUTER PROGRAMMER	61 SPEECH PATHOLOGIST	59 SOCIAL SCIENTIST
DOT	045.088+.108	020.188	079.1 8	050.-059.
OOH	563-565	109-113	514-516	529-541

YOUR LOWEST SCORES APPEARED ON THE FOLLOWING SCALES AND INDICATE THE GREATEST DISSIMILARITY BETWEEN YOUR ANSWERS AND THOSE OF MALES IN THESE OCCUPATIONS--

	-6 AGRIBUSINESS MANAGER	-2 BUYER	0 FUNERAL DIRECTOR	2 LIFE INSURANCE AGENT
DOT	409.168	162.158	187.168	250.258
OOH	627	150-151	173-175	226-228

* *

THE U.S. GOVERNMENT PRINTING OFFICE ALSO HAS A REPRINT SERIES THAT WILL PROVIDE YOU WITH ADDITIONAL INFORMATION ON EMPLOYMENT, WORKING CONDITIONS, AND PAY FOR MANY OCCUPATIONS. REPRINT BULLETINS ARE SUBSECTIONS OF THE OOH IN PAMPHLET FORM. LISTED BELOW IS THE BULLETIN NUMBER AND COST FOR EACH REPRINT THAT WOULD BE IMPORTANT FOR YOU IF THE OOH IS NOT READILY AVAILABLE.

BULLETIN NO.	COST	TITLE
1[illegible]5	$.25	PSYCHOLOGISTS
2[illegible]	$.25	COMPUTER + RELATED OCCUPATIONS
95	$.25	THERAPY AND REHABILITATION OCCUPATIONS
1[illegible]1	$.30	ANTHROPOLOGISTS, ECONOMISTS, SOCIOLOGISTS.

YOUR HIGH OCCUPATIONAL SCALE SCORES SHOULD BE VIEWED AS IMPORTANT INDICATORS OF HOW YOUR INTERESTS MATCH THOSE OF EMPLOYED WORKERS. YOU SHOULD USE THESE RESULTS AS GUIDES AND SEARCH OUT ADDITIONAL INFORMATION ABOUT THOSE AREAS WHERE YOUR INTERESTS AND ABILITIES ARE FOCUSED.

ASK YOUR LIBRARIAN FOR FURTHER INFORMATION ON JOBS IN THESE AREAS, TALK TO PEOPLE IN THESE FIELDS, READ BOOKS AND PERIODICALS THAT ARE RELEVANT, AND SO FORTH.

TO LEARN MORE ABOUT THE LIFE AND WORK OF PEOPLE IN OCCUPATIONS WHERE YOU HAVE HIGH SCORES, THE FOLLOWING BOOKS ARE SUGGESTED FOR FURTHER INFORMATION--

--WATSON, ROBERT IRVING. GREAT PSYCHOLOGISTS. LIPPINCOTT, 1968.
--LUFTIG, MILTON. COMPUTER PROGRAMMER. ARCO, 1967.
--BLACK, MARTHA E. SPEECH CORRECTION IN THE SCHOOLS. PRENTICE, 1964.
--RANDOM, MATTHEW. THE SOCIAL SCIENTISTS IN AMERICAN INDUSTRY. RUTGERS UNIVERSITY PRESS, 1970.

TO LEARN MORE ABOUT THE DAILY ROUTINES OF PEOPLE IN OCCUPATIONS WHERE YOU HAVE HIGH SCORES, YOU MAY WISH TO CONSIDER SPENDING SOME TIME TRYING THESE ACTIVITIES--

--VISIT A MENTAL HEALTH CLINIC OR VOLUNTEER FOR AN EMERGENCY SERVICE TELEPHONE LINE.
--VISIT A COMPUTER TRAINING SCHOOL OR SPEND A DAY WITH A COMPUTER PROGRAMMER.
--SPEND A DAY AT A SPEECH PATHOLOGISTS OFFICE.
--SPEND SOME TIME WITH A SOCIAL WORKER OR DO SOCIAL SERVICE VOLUNTEER WORK.

****

YOU SHOULD PAY PARTICULAR ATTENTION TO YOUR HIGH AND LOW SCORES ON ALL THREE SETS OF SCALES THAT HAVE BEEN PROVIDED -- OCCUPATIONAL THEMES, BASIC INTEREST SCALES, AND OCCUPATIONAL SCALES. YOU MAY FIND THAT EVEN THOUGH YOU MAY HAVE A HIGH SCORE IN ONE OF THE BASIC AREAS SUCH AS ART, YOUR RESPONSES MAY HAVE INDICATED THAT YOU WOULD NOT LIKE THE EVERYDAY ROUTINE AND LIFE SITUATIONS OF AN ARTIST AND YOU HAD A LOW SCORE ON THE ARTIST SCALE. SUCH COMPARISONS CAN HELP YOU BETTER EVALUATE THE TYPES OF WORK SITUATIONS THAT MATCH MOST CLOSELY YOUR INTEREST AND WOULD BE MOST SATISFYING TO YOU.

* * *

THE PRECEDING HAS BEEN AN ANALYSIS OF YOUR INTERESTS ON THE OCCUPATIONAL SCALES. THESE SCALES ALSO ARE RELATED TO THE PREVIOUS 6 GENERAL OCCUPATIONAL THEME SCALES AND BASIC INTEREST SCALES. SUMMARIZED BELOW ARE YOUR SCORES ON EACH THEME SCALE, RELATED BASIC INTEREST SCALES, AND HIGHEST OCCUPATIONAL SCALES WITH EACH AREA.

REMEMBER, IN A GENERAL POPULATION, THE AVERAGE ADULT HAS A SCORE OF 50 ON EACH OF THE 6 THEME SCALES AND 23 BASIC INTEREST SCALES. HOWEVER, ON THE OCCUPATIONAL SCALES, SCORES OF 50 INDICATE SIMILARITY WITH THE AVERAGE WORKER IN THE OCCUPATION AND SCORES OF 26-44 ARE AVERAGE FOR PEOPLE-IN-GENERAL.

I. 69 I-THEME ---
(INVESTIGATIVE)

69 SCIENCE
67 MATHEMATICS
63 MEDICAL SCIENCE
56 MEDICAL SERVICE

64 PSYCHOLOGIST
62 COMPUTER PROGRAMMER
61 SPEECH PATHOLOGIST
59 SOCIAL SCIENTIST

II. 61 A-THEME ---
(ARTISTIC)

56 ART
55 WRITING
54 MUSIC/DRAMA

55 MUSICIAN
54 ARCHITECT
50 PHOTOGRAPHER
48 ARTIST

III. 55 R-THEME ---
(REALISTIC)

59 MECHANICAL
58 AGRICULTURE
58 NATURE
53 ADVENTURE
41 MILITARY ACTIVITIES

49 CARTOGRAPHER
46 VETERINARIAN
45 NAVY OFFICER
42 MERCHANT MARINE OFF.

IV. 48 C-THEME ---
(CONVENTIONAL)

44 OFFICE PRACTICES

13 BUSINESS EDUC. TCH.
8 ACCOUNTANT
7 BANKER

V. 42 S-THEME ---
(SOCIAL)

52 TEACHING
44 DOMESTIC ARTS
43 SOCIAL SERVICE
39 ATHLETICS
35 RELIGIOUS ACTIVITIES

44 ELEMENTARY TEACHER
34 PHYSICAL THERAPIST
29 PUBLIC ADMINISTRATOR
27 LICENSED PRAC. NURSE
27 PERSONNEL DIRECTOR

VI. 38 E-THEME ---
(ENTERPRISING)

49 PUBLIC SPEAKING
45 LAW/POLITICS
41 BUSINESS MANAGEMENT
37 SALES
34 MERCHANDISING

37 INVESTMENT FUND MGR.
30 CHIROPRACTOR
27 PHARMACIST
26 LAWYER

* * *

THE FOLLOWING ARE YOUR SCORES ON OCCUPATIONAL SCALES DEVELOPED ON THE OPPOSITE SEX. ALTHOUGH OCCUPATIONAL SCALES DEVELOPED ON THE SEX CORRESPONDING TO YOURS ARE MORE LIKELY TO BE BETTER PREDICTORS FOR YOU THAN SCALES FOR THE OTHER SEX WOULD BE, THE RESULTS ARE PRESENTED AS EXTRA INFORMATION AND SO AS TO NOT LIMIT YOUR CHOICE OF A CAREER NOT REPRESENTED BY YOUR SEX.

-----YOUR SCORES ON SCALES DEVELOPED ON FEMALE OCCUPATIONAL SAMPLES-----

64 COMPUTER PROGRAMMER	63 SPEECH PATHOLOGIST	63 PSYCHOLOGIST	62 CHEMIST
62 PHYSICIST	61 PHYSICIAN	6 MATHEMATICIAN	60 LANGUAGE INTERPRETER
58 COLLEGE PROFESSOR	57 MEDICAL TECHNOLOGIST	56 ENGINEER	55 VETERINARIAN
54 DENTIST	51 PHARMACIST	5 OPTOMETRIST	49 OCCUP. THERAPIST
46 PHYSICAL THERAPIST	46 MATH-SCIENCE TEACHER	43 ADVERTISING EXEC.	42 LIBRARIAN
39 X-RAY TECHNICIAN	38 ARMY OFFICER	38 ARTIST	37 MUSICIAN
36 DENTAL HYGIENIST	36 REPORTER	35 REGISTERED NURSE	34 ENTERTAINER
34 LAWYER	31 DIETITIAN	31 ART TEACHER	31 SOCIAL SCIENCE TCHR.
3 ACCOUNTANT	29 ENGLISH TEACHER	29 YWCA STAFF	25 RECREATION LEADER
25 GUIDANCE COUNSELOR	23 LANGUAGE TEACHER	22 BANKER	21 INSTRUMENT ASSEMBLER
21 PHYS. EDUCATION TCHR	21 SOCIAL WORKER	2 LICENSED PRAC. NURSE	19 ELEMENTARY TEACHER
18 DENTAL ASSISTANT	17 INTERIOR DECORATOR	15 CHRISTIAN EDUC DIR.	15 CREDIT MANAGER
14 LIFE INSURANCE AGENT	12 BEAUTICIAN	1 FLIGHT ATTENDANT	8 BUYER
7 SECRETARY	6 EXECUTIVE HOUSEKEEPE	5 BUSINESS EDUC. TCHR.	3 HOME ECONOMICS TCHR.
3 DEPT. STORE SALES			

THE FOLLOWING ARE THE ADMINISTRATIVE INDICES AND SPECIAL SCALES. DATA FOR I-VII ARE YOUR LIKE, INDIFFERENT, AND DISLIKE RESPONSE PERCENTAGES (LP-IP-DP) TO THE SEVEN SECTIONS OF THE INVENTORY. TOTAL RESPONSES AND INFREQUENT RESPONSES ARE INTERNAL CHECKS AND THE RESULTS FOR THE SPECIAL SCALES ARE SUMMARIZED BELOW.

		RESPONSE PERCENTAGES		
		LP	IP	DP
I.	OCCUPATIONS	33	27	40
II.	SCHOOL SUBJECTS	53	14	33
III.	ACTIVITIES	33	24	41
IV.	AMUSEMENTS	36	36	28
V.	TYPES OF PEOPLE	33	42	25
VI.	PREFERENCES	50	13	37
VII.	CHARACTERISTICS	64	29	7

TOTAL RESPONSES = 323

INFREQUENT RESPONSES = 11

AOR = 69

IE = 6

YOUR SCORE ON THE ACADEMIC INTEREST SCALE (AOR) INDICATES A SIMILARITY OF YOUR INTERESTS WITH PEOPLE WHO GO TO COLLEGE AND OBTAIN AN ADVANCED DEGREE BEYOND THEIR BACCALAUREATE DEGREE SUCH AS DOCTORS, SCIENTISTS, AND OTHER PROFESSIONALS.

YOUR HIGH SCORE ON THE INTROVERSION-EXTROVERSION (IE) SCALE INDICATES YOU PREFER WORKING WITH THINGS RATHER THAN PEOPLE. SKILLED TRADES WORKERS, SCIENTISTS, AND ARTISTS SCORE HIGH ON THIS SCALE WHILE SALES AND SERVICE OCCUPATIONS HAVE LOW SCORES ON THIS SCALE.

```
***************************************
*****                             *****
*****      CLOSING COMMENTS       *****
*****                             *****
***************************************
```

YOU HAVE BEEN SCORED ON A BROAD RANGE OF GENERAL INTERESTS AND SPECIFIC OCCUPATIONS. YOU SHOULD NOT BE TOTALLY SET ON ANY ONE PARTICULAR OCCUPATION WHERE YOUR SCORE IS HIGH, AT LEAST NOT AT AN EARLY AGE. IN THE WORLD OF WORK THERE ARE HUNDREDS OF SPECIALITIES AND PROFESSIONS, AND YOU SHOULD USE THIS INFORMATION AS A GUIDE FOR FURTHER EXPLORATION.

THE RESULTS PRESENTED FOR YOU HAVE BEEN BASED ON YEARS OF RESEARCH ON INTEREST PREFERENCES AND ANALYSES OF HUNDREDS OF THOUSANDS OF PEOPLE. STUDIES MADE YEARS LATER OF EMPLOYED PEOPLE WHO COMPLETED EARLIER VERSIONS OF THIS FORM IN THEIR HIGH SCHOOL OR COLLEGE DAYS HAVE SHOWN THAT ABOUT ONE-HALF END UP IN OCCUPATIONS COMPATIBLE WITH THEIR SCORES, AND MOST OF THEM LIKE THEIR WORK. AMONG THOSE WHO END UP IN OCCUPATIONS NOT COMPATIBLE WITH THEIR RESULTS, MANY SAY THEY DO NOT LIKE THEIR WORK, OR THEY ARE DOING THE JOB IN SOME UNUSUAL MANNER.

FURTHER RESEARCH HAS SHOWN THAT SCORES ON INTEREST INVENTORIES ARE VERY STABLE. ADULTS WILL SHOW VERY LITTLE CHANGE IN THEIR SCORES IF TESTED OVER A NUMBER OF YEARS. HIGH SCHOOL STUDENTS WILL SHOW SOME CHANGE IF TESTED AND RETESTED AFTER TWO OR THREE YEARS AND MORE CHANGE AFTER TEN OR FIFTEEN YEARS. GENERALLY, SCORES ON INTEREST INVENTORIES ARE VERY STABLE.

EACH PERSON IS UNIQUE AND NO TEST CAN PREDICT WITH PERFECT ACCURACY ALL THE DIVERSITY AMONG VARIOUS INDIVIDUALS. THUS, THE RESULTS PRESENTED TO YOU SHOULD BE USED AS GUIDELINES IN HELPING YOU TO BETTER UNDERSTAND YOUR CAREER INTERESTS.

FOR ADDITIONAL INFORMATION, THE FOLLOWING SOURCES ARE RECOMMENDED--
--CONSULT A PROFESSIONALLY TRAINED COUNSELOR
--THE HANDBOOK FOR THE STRONG VOCATIONAL INTEREST BLANK ($20.00) WRITE STANFORD UNIVERSITY PRESS
--THE MANUAL FOR THE STRONG-CAMPBELL INTEREST INVENTORY ($5.00) WRITE NATIONAL COMPUTER SYSTEMS.

091174 DR. CHARLES B. JOHANSSON, NATIONAL COMPUTER SYSTEMS (NCS), 4401 W. 76TH STREET, MINNEAPOLIS, MINNESOTA 55435. ADDITIONAL TESTING MATERIALS MAY BE PURCHASED FROM NCS.

COPYRIGHT 1933 (RENEWED 1961), 1938 (RENEWED 1965), 1945 (RENEWED 1974), 1959, 1964, 1965, 1966, 1968, 1969, 1972, 1974 BY THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY. ALL RIGHTS RESERVED. PRINTED UNDER LICENSE FROM STANFORD UNIVERSITY PRESS, STANFORD, CALIFORNIA 94305.

******END OF REPORT FOR NO NAME ****** 1 /16/74 999 0025

109

CLEMANS' LAMBDA-BASED SCALES OF THE NAVY VOCATIONAL INTEREST INVENTORY

Joyce E. Dann

Navy Personnel Research and Development Center
San Diego, California 92152

The Navy has long been concerned with the problem of optimally assigning enlisted personnel to jobs or ratings. Under the present system, a recruit's assignment is based primarily upon his scores on ability measures, his work experience, and the Navy's manpower needs. Job preferences are also considered but are often difficult to assess because the individual may have limited knowledge of his vocational interests and of the way in which his interests relate to the complex Navy job structure.

The purpose of this paper is to describe the application of a vocational interest inventory--the Navy Vocational Interest Inventory (NVII)--to the recruit assignment problem. The inventory was developed by Kenneth E. Clark in 1951 (Clark, 1961), and was later shortened and published commercially as the Minnesota Vocational Interest Inventory. The Navy version (or NVII) is composed of 190 forced-choice item triads. Each triad consists of three job tasks or activities from among which the individual selects the task he likes most and the task he likes least.

Based on the NVII items, Clark constructed 19 Navy occupational-type scales using Strong's men-in-general method (Strong, 1943) and nine homogeneous or area scales. However, neither type of scale was used extensively.

Recently, the Navy has revived its interest in vocational interest measures, and the scales developed by Clark have been re-evaluated using more current samples of enlisted personnel. In several studies at the Navy Personnel Research and Development Center, San Diego, both the occupational and area-type scales were found to differentiate well among men in different Navy schools (Abrahams, Lau, & Neumann, 1968; Lau & Abrahams, 1969) and to predict such criteria as Navy Class "A" school achievement (Abrahams, Lau, & Neumann, 1968) and reenlistment (Lau, Lacey, & Abrahams, 1969). The scales also showed substantial test-retest reliabilities for reenlistees (Lau & Abrahams, 1970) and non-reenlistees (Lau & Abrahams, 1971) over four to six years.

Based upon these favorable results, the Navy began a large-scale effort in 1971 to update the NVII scales using more recent samples of enlisted men. Fifteen ratings were selected for the initial scale-construction effort, with additional scales planned for future construction. Some of the ratings were chosen to be representative of broad clusters of ratings such as mechanical, medical, food service, electrical-electronics, and clerical fields, so that an individual's scores could be generalized to indicate his interest in other ratings similar to the 15. Other ratings were chosen for their similarity to each other to permit evaluation of the scales' ability to discriminate between closely related ratings. The scale-construction group for each rating included between 105 and 361 men who were experienced in the Navy, satisfied with their rating, and doing work typical for that rating.

A scale-construction technique developed by Clemans (1958) was selected over the more traditional men-in-general method because Kuder found that it discriminated better than the men-in-general method between occupations of his Occupational Interest Survey (Kuder, 1968). Clemans' method also circumvented the unresolved methodological problems in constructing a men-in-general group. Briefly, the men-in-general method contrasts the interest inventory responses of members of an occupational group with responses of a general reference group. An individual's score thus reflects the extent to which he differs from the reference group in the same way as do members of the occupation. By contrast, Clemans' scale-construction technique yields scores, or lambda coefficients, based upon a direct correlation between the individual's inventory responses and responses of members of an occupation.

To construct lambda scales, the proportion of each rating criterion group endorsing each of the NVII's 1,140 possible response alternatives (i.e., one "like" and one "dislike" alternative for each of the three tasks comprising each of the NVII's 190 items) was determined. These proportions for each rating criterion group constituted the continuous variable for computing a point-biserial correlation, while the individual's selection or non-selection of a response alternative constituted the dichotomous variable. The correlation was divided by the maximum like-signed point-biserial correlation which the individual could have attained with the criterion group, and a positive or negative sign was affixed to the result according to the sign of the individual's actual correlation with the criterion group. The resulting lambda scores could assume values between -1.00 and 1.00. (For a more complete discussion of the development of lambda scales of the NVII, see Dann and Abrahams, 1973.)

The remainder of this paper will consider the lambda scales' intercorrelations, test-retest reliability, and validity.

Table 1 shows a lambda scale intercorrelation matrix based upon a sample of 1,055 recruits tested in 1971.

As the table indicates, some of the more unique ratings, such as Hospital Corpsman and Storekeeper, had low correlations with all of the other ratings. By contrast, some closely related ratings, such as Sonar Technician and Electronics Technician, showed correlations with each other as high as .99. In general, these correlations were considerably higher than expected and than those reported for the Strong Vocational Interest Blank (SVIB) (Campbell, 1971), the Kuder Occupational Interest Survey (Kuder, 1968), or for Clark's NVII occupational scales (Clark, 1961). The disparity between the lambda NVII scales and both Clark's NVII scales and the SVIB scales probably reflects differences in scoring methodologies. That is, the men-in-general method used in constructing Clark's scales and the SVIB scales isolates interests unique to a particular occupation, while the lambda method taps both unique interests and interests shared by many occupational groups. The common core of interests tapped using the lambda method would naturally tend to increase the magnitude of scale intercorrelations.

The difference in intercorrelation magnitudes between the Kuder and NVII lambda scales probably reflects the method of selecting the occupational groups in each case. Kuder based his correlations upon 23 "core groups" selected for their diversity and representativeness of the different types of occupations on his profile. The lambda scale intercorrelations are derived from some ratings chosen for their diversity, but also from others selected specifically for their similarity to each other. These similar ratings should be highly intercorrelated. However, the high degree of intercorrelation in this case may limit the discriminating ability of the scales.

Figure 1 shows test-retest reliabilities on the lambda scales for men administered the NVII in 1964 or 1965 prior to entering one of seven Navy Class "A" schools and readministered the inventory four to six years later. The reliabilities were determined by rank-ordering each individual's test and retest scores and computing Spearman's rank difference correlations between the profiles from

TABLE 1

Intercorrelations of 1,055 Recruits-in-General on 15 Lambda Scales

Scale	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
(1) Quartermaster		.89	.88	.91	.88	.14	.66	.77	.78	.82	.80	.85	.99	.85	.33
(2) Sonar Technician	.89		.99	.95	.81	-.12	.63	.93	.93	.97	.92	.95	.91	.99	.14
(3) Electronics Technician	.88	.99		.94	.80	-.16	.60	.91	.92	.97	.91	.94	.90	.98	.15
(4) Radioman	.91	.95	.94		.91	.13	.58	.86	.88	.92	.86	.90	.92	.94	.07
(5) Data Processing Technician	.88	.81	.80	.91		.44	.56	.72	.74	.77	.73	.78	.87	.79	.12
(6) Storekeeper	.14	-.12	-.16	.13	.44		.09	-.12	-.10	-.13	-.09	-.07	.08	-.13	-.14
(7) Commissaryman	.66	.63	.60	.58	.56	.09		.70	.68	.66	.75	.74	.67	.65	.12
(8) Engineman	.77	.93	.91	.86	.72	-.12	.70		.99	.98	.99	.99	.78	.97	-.12
(9) Boiler Technician	.78	.93	.92	.88	.74	-.10	.68	.99		.99	.99	.99	.79	.97	-.12
10) Electrician's Mate	.82	.97	.97	.92	.77	-.13	.66	.98	.99		.97	.98	.83	.99	-.03
(11) Equipment Operator	.80	.92	.91	.86	.73	-.09	.75	.99	.99	.97		.99	.81	.96	-.08
(12) Aviation Ordnanceman	.85	.95	.94	.90	.78	-.07	.74	.99	.99	.98	.99		.86	.98	-.03
(13) Air Controlman	.99	.91	.90	.92	.87	.08	.67	.78	.79	.83	.81	.86		.87	.34
(14) Aviation Electrician's Mate	.85	.99	.98	.94	.79	-.13	.65	.97	.97	.99	.96	.98	.87		.04
(15) Hospital Corpsman	.33	.14	.15	.07	.12	-.14	.12	-.12	-.12	-.03	-.08	-.03	.34	.04	

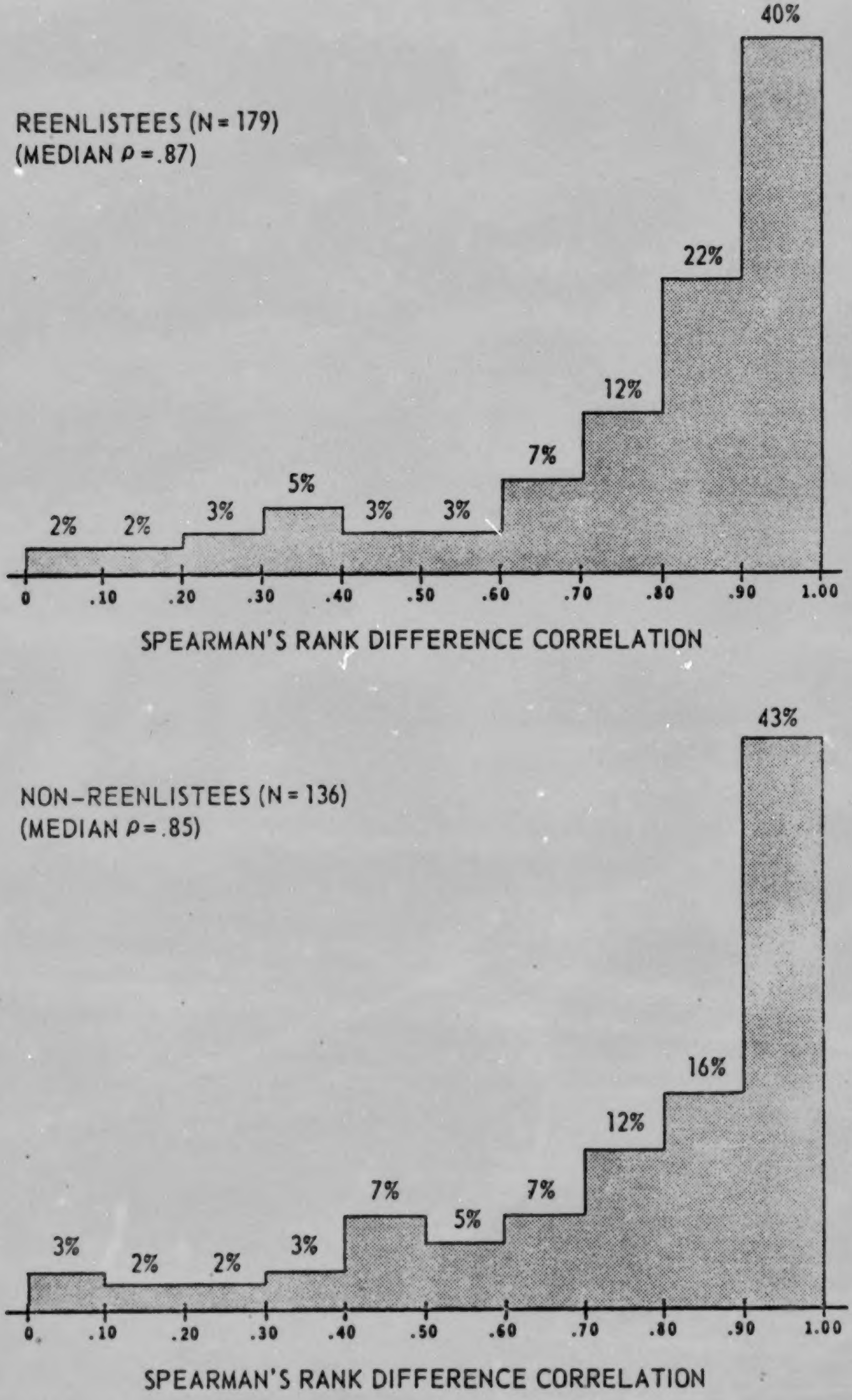


Fig. 1. Distribution of intra-individual test-retest correlations for reenlistees and non-reenlistees.

the two occasions. Separate distributions of correlations are shown for individuals who reenlisted in the Navy and for individuals who left the Navy after their first term of enlistment.

Among reenlistees, the median correlation was .87, with only 15 percent of the correlations .50 or less and 40 percent above .90. Among non-reenlistees, the median correlation was .85. Approximately 17 percent of the group's correlations were .50 or smaller, while 43 percent were above .90. Thus, reenlistees and non-reenlistees were highly similar to each other in their reliability distributions and exhibited substantial stability of lambda score profiles over the time periods considered. Reliabilities were roughly comparable to those obtained by Kuder for the lambda scales of his Occupational Interest Survey (Kuder, 1968).

To be useful in recruit classification, however, the lambda NVII scales must also be related to criteria such as job satisfaction. That is, satisfied individuals, more often than dissatisfied individuals, should exhibit interests characteristic of their rating.

To investigate this relationship, a sample of men who were experienced in, and doing work typical for, their rating was used. The individuals were divided into three job satisfaction subgroups. The first subgroup consisted of men who indicated they would choose the same rating again if given the chance. The second group indicated they would choose a rating very similar to their present one, and the third that they would select a rating in a completely different occupational field.

The percentage of men in each of these satisfaction subgroups who had interests consonant with their rating and who would thus have been directed to that rating was then determined. Results are shown in Table 2. In computing the percentages, individuals were considered to have interests consonant with those ratings for which they received their highest score or a score within the .01 standard error of measurement confidence interval of their highest score.

Among individuals who would select the same rating, a median of 76 percent (with a range from 52 to 93 percent) had interests consonant with that rating. This contrasts with a median of only 45 percent of individuals selecting a completely different rating who had the interests of their rating group. For most ratings,

TABLE 2

Percentage of Men in Each Rating Who Would Choose the Same, a Similar, and a Different Rating Who Obtained Their Highest Score or a Score Within .06* of Their Highest on the Scale for Their Own Rating

Rating	Sample Sizes			Percentage of Men Obtaining Highest Score or Score Within .06 of Highest On Scale			Difference in Percentages Between Men Who Would Choose the Same Versus a Different Rating
	Choose Same Rating	Choose Similar Rating	Choose Different Rating	Choose Same Rating	Choose Similar Rating	Choose Different Rating	
Quartermaster	71	68	249	52	44	50	2
Sonar Technician	68	223	136	76	75	36	40
Electronics Technician	124	157	86	83	75	44	39
Radioman	69	224	157	54	50	29	25
Data Processing Technician	180	47	71	57	49	24	33
Storekeeper	63	151	165	63	68	29	34
Commissaryman	79	56	276	73	30	35	38
Engineman	71	152	103	93	89	56	37
Boiler Technician	29	178	223	86	89	59	27
Electrician's Mate	72	202	133	89	73	47	42
Equipment Operator	91	38	43	79	87	49	30
Aviation Ordnanceman	53	155	281	74	58	49	25
Air Controlman	180	27	76	56	48	43	13
Aviation Electrician's Mate	100	214	96	86	73	48	38
Hospital Corpsman	126	44	78	92	82	45	47
Median Percentages				76	73	45	

*.06 delineates the .01 standard error of measurement confidence interval.

over 30 percent more satisfied than dissatisfied individuals showed consonant interests.

Individuals who would select a similar rating generally achieved percentages intermediate between those of the more extreme satisfaction groups.

The relationship between NVII interests and job satisfaction may also be evaluated by comparing the mean ranks of men who would choose the same, a similar, or a different rating on the scale for their own rating.

Table 3 provides this information and indicates that the more satisfied individuals generally scored higher on their rating's scale than did the less satisfied individuals. In no case was the mean for men who would choose the same rating lower than that for individuals preferring a completely different field. For most ratings, the difference in means exceeded half the larger standard deviation for the two extreme satisfaction groups.

Both of the preceding analyses, then, indicate a substantial relationship between an individual's job satisfaction and his NVII scores.

For individuals who are satisfied in a rating, several further questions may be raised: Do individuals who prefer to remain in their rating score higher on the scale for that rating than on scales for other ratings? And, do satisfied men in a rating score higher on their relevant scale than satisfied men in other ratings score on that scale?

Table 4 provides information for answering these questions based on a cross-validation sample of men who met the same job experience, job satisfaction, and typicality of work criteria used to delineate the original scale-construction group.

As the figures indicate, most ratings--with the exception of Electronics Technician and Storekeeper--did obtain their highest mean rank on the scale for their own rating.

Most ratings also obtained a higher mean rank than any other rating on their own scale, with only seven exceptions to this trend among the 210 off-diagonal entries in Table 4.

To summarize the results presented in this paper, then:

TABLE 3

Mean Ranks of Individuals in Each Rating Who Would Choose the Same, a Similar, and a Different Rating on the Scale for Their Own Rating

Present Rating	Sample Sizes			Mean Ranks			Standard Deviations of Ranks		
	Choose Same Rating	Choose Similar Rating	Choose Different Rating	Choose Same Rating	Choose Similar Rating	Choose Different Rating	Choose Same Rating	Choose Similar Rating	Choose Different Rating
Quartermaster	71	68	249	4.4	3.5	4.4	3.5	2.5	3.4
Sonar Technician	68	223	136	3.4	3.6	5.7	2.3	2.0	2.7
Electronics Technician	124	157	86	3.1	3.5	5.5	2.3	2.7	3.6
Radioman	69	224	157	4.1	4.3	6.3	3.1	2.8	3.2
Data Processing Technician	180	47	71	3.9	5.1	7.8	3.6	4.2	4.0
Storekeeper	63	151	165	3.9	4.4	9.3	4.2	5.3	5.8
Commissaryman	79	56	276	3.1	5.8	7.3	3.4	4.7	4.8
Engineman	71	152	103	2.7	3.2	6.6	2.5	3.1	4.9
Boiler Technician	29	178	223	3.9	4.6	6.4	3.9	2.8	4.6
Electrician's Mate	72	202	133	3.1	4.0	6.5	2.5	3.0	3.8
Equipment Operator	91	38	43	3.7	3.6	6.8	3.5	2.2	3.5
Aviation Ordnanceman	53	155	281	4.4	5.5	5.8	2.6	3.0	2.9
Air Controlman	180	27	76	4.3	4.5	5.2	3.5	3.6	3.8
Aviation Electrician's Mate	100	214	96	3.4	4.1	6.3	2.2	2.5	3.1
Hospital Corpsman	126	44	78	1.8	2.3	7.0	2.6	3.2	5.9

TABLE 4

Mean Ranks of Men in Each of 15 Ratings on Each Lambda Scale

Present Rating	Lambda Scale														
	Quartermaster	Sonar Technician	Electronics Technician	Radioman	Data Processing Technician	Storekeeper	Commissaryman	Engineman	Boiler Technician	Electrician's Mate	Equipment Operator	Aviation Ordnanceman	Air Controlman	Aviation Electrician's Mate	Hospital Corpsman
Quartermaster	4.4	6.7	8.3	6.2	6.9	9.6	9.7	10.3	9.7	9.8	8.8	6.2	5.1	8.6	9.9
Sonar Technician	7.6	3.4	4.2	6.4	9.0	13.7	12.3	8.9	9.1	5.2	9.6	6.8	7.8	4.2	11.8
Electronics Technician	8.1	2.9	3.1	6.5	9.0	14.0	12.9	8.8	8.6	4.4	9.9	7.1	8.4	3.9	12.4
Radioman	6.7	4.9	6.7	4.1	6.8	10.5	11.0	10.6	9.7	7.2	10.1	7.2	6.5	6.2	11.9
Data Processing Technician	6.0	6.5	8.2	6.4	3.9	8.5	10.0	10.5	9.7	8.1	10.1	6.9	6.4	7.9	10.9
Storekeeper	5.1	8.4	11.2	5.5	3.3	3.9	5.9	12.1	10.7	10.5	9.6	7.1	5.8	10.8	10.2
Commissaryman	6.6	9.0	10.4	9.5	8.0	9.0	3.1	8.6	9.0	9.0	6.3	5.4	7.2	9.1	9.9
Engineman	9.5	7.4	8.4	9.1	11.0	13.8	11.3	2.7	3.2	4.8	4.1	4.1	10.3	5.8	14.4
Boiler Technician	9.3	6.7	7.6	8.5	10.5	13.6	11.9	4.1	3.5	5.2	5.4	4.5	9.7	5.8	13.8
Electrician's Mate	9.5	4.9	5.9	7.8	10.7	14.2	12.6	5.7	6.1	3.1	7.2	5.4	10.0	3.6	13.4
Equipment Operator	8.1	8.0	9.7	8.0	9.6	12.4	10.4	5.4	5.3	6.6	3.7	4.0	8.2	7.1	13.5
Aviation Ordnanceman	7.7	7.5	8.6	8.0	9.2	12.1	9.8	6.4	6.2	6.8	5.8	4.4	8.2	6.7	12.5
Air Controlman	4.9	6.5	7.5	6.4	7.2	10.9	10.4	10.1	9.9	9.1	8.5	6.1	4.3	7.7	10.3
Aviation Electrician's Mate	9.2	4.3	5.6	6.7	9.6	14.0	12.7	7.1	7.2	3.9	8.2	6.0	8.9	3.4	13.2
Hospital Corpsman	3.7	6.6	6.8	8.1	6.8	9.7	7.2	12.7	13.1	10.8	10.8	9.0	4.2	8.7	1.8

(1) The lambda scales of the NVII were more highly intercorrelated than men-in-general scales of the SVIB or NVII or than the 23 core group lambda scales of Kuder's Occupational Interest Survey.

(2) Profiles of scores based upon lambda scales of the NVII remained quite stable over intervals of at least four to six years.

(3) Most satisfied men in a rating would have been directed to that rating based upon their NVII scores, while much smaller percentages of dissatisfied individuals in a rating would have been directed toward that rating.

(4) Men who would choose the same rating again obtained higher mean ranks on the scale for their rating than did men in that rating who would choose a similar or different one.

(5) Satisfied men in a rating obtained higher mean ranks on the scale for their own rating than on scales for other ratings. They also achieved higher mean ranks on their scale than did satisfied men from other ratings.

REFERENCES

Abrahams, N. M., Lau, A. W., & Neumann, I. *An analysis of the Navy Vocational Interest Inventory as a predictor of school performance and rating assignment*. San Diego: U. S. Naval Personnel Research Activity, October 1968. (Research Report SRR 69-11)

Campbell, D. P. *Handbook for the Strong Vocational Interest Blank*. Stanford University Press, 1971.

Clark, K. E. *The vocational interests of nonprofessional men*. Minneapolis: University of Minnesota Press, 1961.

Clemans, W. V. An index of item-criterion relationship. *Educational and Psychological Measurement*, 1958, 18, 167-172.

Dann, J. E., & Abrahams, N. M. *Occupational scales of the Navy Vocational Interest Inventory: I. Development*. San Diego: Navy Personnel Research and Development Center, October 1973. (Technical Report NPRDC TR 74-4)

Kuder, G. F. *Kuder Occupational Interest Survey general manual*. Chicago: Science Research Associates, Inc., 1968.

Lau, A. W., Lacey, L. A., & Abrahams, N. M. *An analysis of the Navy Vocational Interest Inventory as a predictor of career motivation*. San Diego: U. S. Naval Personnel Research Activity, June 1969. (Research Report SRR 69-27)

Lau, A. W., & Abrahams, N. M. *Area scales of the Navy Vocational Interest Inventory as predictors of school performance and rating assignment*. San Diego: U. S. Naval Personnel Research Activity, July 1969. (Research Report SRR 70-1)

Lau, A. W., & Abrahams, N. M. *The stability of scores on the Navy Vocational Interest Inventory*. San Diego: Naval Personnel and Training Research Laboratory, March 1970. (Research Report SRR 70-17)

Lau, A. W., & Abrahams, N. M. *Reliability and predictive validity of the Navy Vocational Interest Inventory*. San Diego: Naval Personnel and Training Research Laboratory, February 1971. (Research Report SRR 71-16)

Strong, E. K. *Vocational interests of men and women*. Stanford University Press, Stanford, 1943.

DEVELOPMENT AND EVALUATION OF STRONG VOCATIONAL INTEREST BLANK SCALES FOR ACADEMIC MAJOR AREAS AT THE U. S. NAVAL ACADEMY[1]

William A. Sands
Navy Personnel Research and Development Center

Robert D. McCullah, LCDR, MSC, USN; PhD
U. S. Naval Academy

BACKGROUND AND PURPOSE

Personnel Utilization

Officer procurement, development, and retention always have been important considerations for the Navy. The advent of an all-volunteer force has placed even greater emphasis upon the importance of effective personnel utilization.

The Navy offers a wide variety of career opportunities. These opportunities must have a sufficiently high payoff to meet career development needs, while simultaneously meeting the operational requirements of the Navy.

Naval Academy Academic Program

The United States Naval Academy (USNA) is an important source of commissioned officers that does offer a variety of career opportunities. However, until 1959, the curriculum was uniform and consisted of a "lock-step" matrix of courses. There was no program involving different majors or minors, and no elective courses were available. In 1959, the validation test technique for advanced placement was introduced for the new plebe class (class of 1963) and an electives program was initiated. This electives program was restricted to academically qualified midshipmen beyond their plebe year. At this point in time, all midshipmen "majored" in Naval Science, and those qualified for the program "minored" in one of 12 specialties.

---

[1] The views expressed herein are those of the authors and do not necessarily reflect those of the U. S. Navy.

The "Majors Program," instituted in 1969, provided majors organized under four broad academic areas: (I) Engineering-Weapons, (II) Mathematics-Science, (III) Humanities, and (IV) Management Science. A recent change in the Majors Program has eliminated the broad area of Management Science (Group IV). The curricula for these management science majors have undergone drastic alteration to include an emphasis on engineering. These revised curricula will be grouped administratively under Group I--Engineering-Weapons.

As indicated in the latest edition of The Majors Program, published by the Academy, there is a marked emphasis on the technical majors of Groups I and II:

> The Naval Academy policy on the selection of majors is clear. Each midshipman selects a major which will meet the needs of the Navy and at the same time be interesting to him. The needs of the Navy take first priority and it has been determined that eighty percent of the Class shall take a technical major, i.e., Group I or II, and twenty percent may choose Group III. Hopefully, the selection of majors by the Class of 1978 will meet the 80/20 quota. If the desired distribution is not obtained by an open, free selection process, steps will be taken to adjust the distribution to meet the Navy goals [p. 9].

The primary purpose of the present research program is the development of tools which will aid the Naval Academy staff in educational-vocational guidance of their students.

Related Research

A rather substantial amount of research on interests and their relationship to various criteria of performance and satisfaction has been accomplished in the past 10-15 years. The Strong Vocational Interest Blank (SVIB) is one of the most useful psychometric instruments for measuring these interests (Campbell, 1966).

Abrahams and Neumann (1973) studied the relationship between SVIB responses and performance in the U. S. Naval Academy. Specifically, these investigators developed scales for predicting three criteria: motivational disenrollment, academic disenrollment, and military aptitude. All these new scales were found to be significantly related to their respective criteria in cross-validation samples.

In another study on the Academy, Neumann and Abrahams (1974) developed an SVIB scale designed to identify USNA applicants with engineering and science interests. The class of 1973 was split into a key-development sample and a cross-validation sample. The criterion employed was dichotomous: Engineering-Science major versus "Other" major. A biserial validity of .57 was obtained in the cross-validation sample. Application of the new SVIB scale to the class of 1976 yielded a biserial validity of .62. The results of this investigation were also reported in a paper presented at a recent Air Force symposium (Abrahams & Neumann, 1974).

Objective

The objective of this research effort is the development of interest scales which will provide an index of the degree of similarity between an individual's measured interests and those of persons who have successfully graduated in each of the three broad major areas of study at the Academy.

PROCEDURE

Instrument

The 1966 edition of the Strong Vocational Interest Blank (SVIB) for men is a paper-and-pencil inventory consisting of 399 items. The content of the items is diverse and includes occupational titles, school subjects, amusements and hobbies, and general occupational activities. The person is asked to respond "Like," "Indifferent," or "Dislike" for most of the items.

Samples

The SVIB was administered on an experimental basis to all midshipmen of the classes of 1971, 1972, and 1973, during plebe summer. Those persons who failed to graduate were removed from the three samples. As previously indicated, there will no longer be a management science major area. Consequently, those persons graduating with a major in the management science area were removed from the three samples. The specific majors in each of the three broad academic areas and the number of graduates in each individual major by class are shown in Table 1.

Criteria

Midshipmen in each of the individual majors were aggregated into the three broad areas, as shown at the bottom of Table 1. The criterion for each scale was dichotomous: membership versus nonmembership in that major group.

TABLE 1

Number of Naval Academy Graduates by Individual Major for the Classes of 1971, 1972, and 1973

Individual Major	Number of Graduates			
	1971	1972	1973	Total
Group I: Engineering-Weapons				
Aerospace Engineering	76	51	38	165
Electrical Engineering	22	24	22	68
General Engineering	0	0	16	16
Marine Engineering	10	13	4	27
Mechanical Engineering	66	41	37	144
Naval Architecture	15	6	10	31
Ocean Engineering	13	18	32	63
Systems Engineering	23	16	12	51
Group II: Mathematics-Science				
Applied Science	12	19	18	49
Chemistry	12	14	35	61
Mathematics	85	58	66	209
Oceanography	95	122	103	320
Operations Analysis	31	51	63	145
Physical Science	0	0	1	1
Physics	29	43	28	100
Group III: Humanities				
American Political Systems	25	49	48	122
International Security Affairs	53	54	56	163
European Studies - French	5	4	7	16
European Studies - German	15	7	3	25
European Studies - Italian	2	1	0	3
Latin American Studies - Spanish	15	12	7	34
Latin American Studies - Portuguese	0	0	2	2
Far Eastern Studies - Chinese	0	3	1	4
Soviet Studies - Russian	2	2	3	7
Economics	23	12	11	46
English	12	6	10	28
History	19	20	30	69

	1971		1972		1973		Total	
Summary	N	%	N	%	N	%	N	%
Group I	225	34	169	26	171	26	565	29
Group II	264	40	307	48	314	47	885	45
Group III	171	26	170	26	178	27	519	26

Scale Construction

The class of 1973 was used to construct the new scales. The approach employed was "empirical criterion keying." This approach involved comparing two groups in terms of their item responses to the SVIB. Those responses which exhibited a specified percentage difference between the two groups were dimensionalized and unit-weighted according to the procedure outlined by Campbell (1971).

The Major Group I Scale was constructed by contrasting the SVIB item responses of engineering-weapons graduates (high criterion group) with the item responses of mathematics-science and humanities graduates combined (low criterion group). Similarly, the Major Group II Scale was constructed by contrasting mathematics-science majors with engineering and humanities majors combined, and the Major Group III Scale contrasted humanities majors with majors in the areas of engineering and mathematics-science combined.

Scale Evaluation

A number of alternative scales were constructed for each major group, varying the percentage difference level required for a response to be included in a scale. The scale which appeared to offer the most promise for each group was chosen for subsequent evaluation. The classes of 1971 and 1972 were used to assess the effectiveness of the scales developed on the class of 1973.

RESULTS

Major Group I Scale

Table 2 presents selected statistical results for the Major Group I Scale, for each Academy class. Cross-validation of this scale on the class of 1971 yielded a .50 biserial correlation.

Another index which is useful in describing the extent to which a predictor variable can separate two groups (e.g., Group I vs. others) is the percent overlap.[2] This index can vary from zero,

---

[2] The percent overlap is based upon the difference in means of the two groups divided by the average of the two standard deviations. This ratio is entered into a table to give the corresponding percent overlap (Tilton, 1937).

TABLE 2

Means, Standard Deviations, Validities, and Percent Overlaps for the Major Group I Scale on Naval Academy Classes of 1971, 1972, and 1973[a]

Class	Criterion Group	Major Group	$\underline{N}$	Mean	Standard Deviation	Validity ($\underline{r}_b$)	Percent Overlap
	High	I	225	119.13	15.82	.50**	66
1971	Low	II+III	435	104.18	17.64		
	Total	I+II+III	660	109.28	18.45		
	High	I	169	119.54	15.15	.44**	68
1972	Low	II+III	477	105.67	18.71		
	Total	I+II+III	646	109.30	18.85		
	High	I	171	120.32	15.33	.56[a]	59
1973	Low	II+III	492	101.68	18.92		
	Total	I+II+III	663	106.49	19.81		

Note.

[a] The statistics for the class of 1973 are biased, as the scale was constructed using these individuals.

**Statistically significant biserial correlations (Alf & Abrahams, 1971):

Class of 1971: $\underline{r}_b \geq .135$, $\underline{p} \leq .01$.

Class of 1972: $\underline{r}_b \geq .142$, $\underline{p} \leq .01$.

representing perfect separation of the two groups, to one hundred, representing a complete overlap of the score distributions for the two groups. Hence, a low percent overlap indicates a highly effective predictor variable. The percent overlap for the class of 1971 on the Group I Scale is 66.

An evaluation of the Major Group I Scale in the class of 1972 shows results which closely parallel those obtained with the class of 1971. Specifically, the cross-validated biserial correlation was .44, while the percent overlap was 68.

The practical significance of applying the Major Group I Scale can be illustrated using an expectancy chart, as shown in Figure 1. The score categories for this and the following five expectancy charts were formed by combining the distributions for the classes of 1971 and 1972. This total distribution was divided into quintiles, each fifth as nearly equal in sample size as possible. Then, treating each class separately, the percentage of Group I majors was computed for each quintile. As indicated in Figure 1, those persons in the class of 1971 scoring in the top quintile on the scale selected Group I majors over five times more frequently than those persons scoring in the lowest quintile. If those persons scoring in the top 60% of the scale distribution were selected, approximately 46% Group I majors would result. This figure may be compared with the base rate[3] of 34%.

Figure 2 shows that results comparable to those obtained for the class of 1971 were found for the class of 1972. Specifically, over six times more Group I majors were observed in the top quintile as were found in the bottom quintile. Selection of those persons scoring in the top 60% of the distribution would result in approximately 37% Group I majors, as compared with a base rate of 26%.

Major Group II Scale

Table 3 presents the results obtained for the Group II Scale, for each class. A biserial correlation of .23 was obtained in cross-validation of the scale on the class of 1971. The percent overlap was 84, reflecting the relatively large overlap between the Group II and the combined Group I and III distributions. The

---

[3] The base rate is the proportion of high criterion group members in the total distribution. For example, in the class of 1971, there were 225 Group I majors out of a total of 660, yielding a base rate of (225/660) .34.

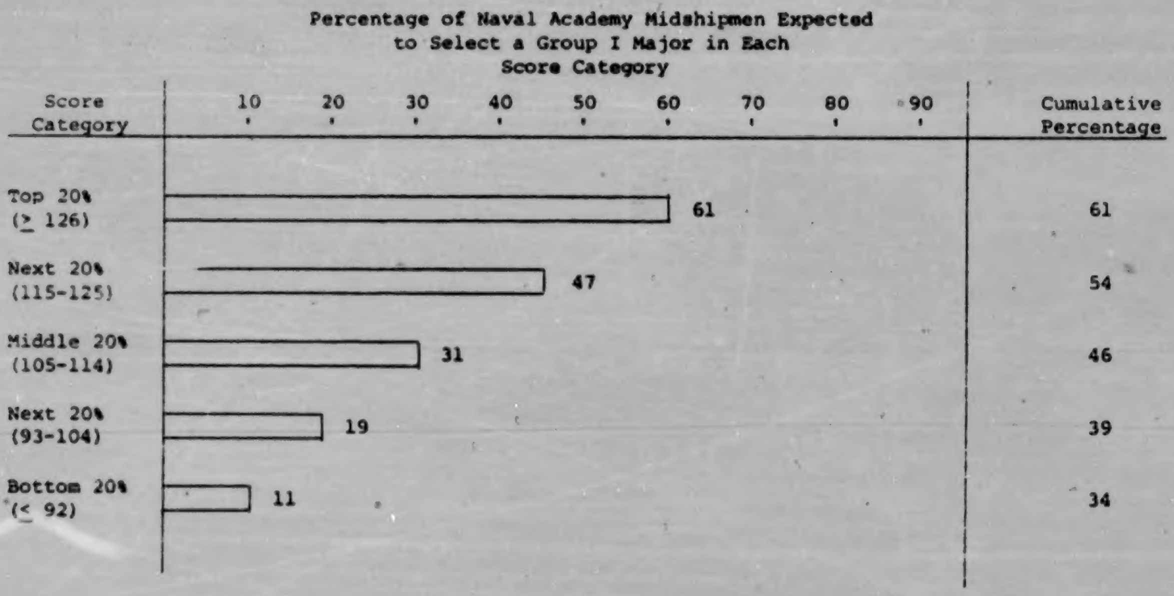


Fig. 1. Expectancy chart showing the relationship between the Major Group I Scale and choice of major for the Naval Academy class of 1971.

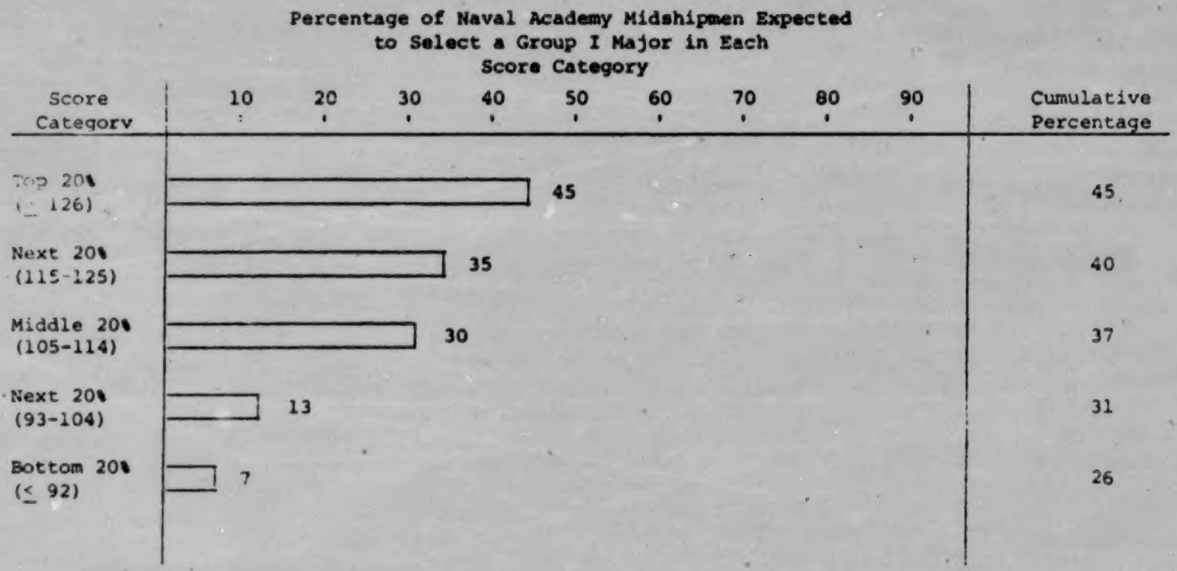


Fig. 2. Expectancy chart showing the relationship between the Major Group I Scale and choice of major for the Naval Academy class of 1972.

TABLE 3

Means, Standard Deviations, Validities, and Percent Overlaps for the Major Group II Scale on Naval Academy Classes of 1971, 1972, and 1973[a]

Class	Criterion Group	Major Group	$\underline{N}$	Mean	Standard Deviation	Validity (r_b)	Percent Overlap
1971	High	II	264	111.89	9.50	.23**	84
	Low	I+III	396	107.69	11.79		
	Total	I+II+III	660	109.37	11.12		
1972	High	II	307	112.99	10.01	.37**	75
	Low	I+III	339	105.94	12.49		
	Total	I+II+III	646	109.29	11.91		
1973	High	II	314	113.24	9.95	.44[a]	70
	Low	I+III	349	104.44	12.90		
	Total	I+II+III	663	108.61	12.40		

Note.

[a] The statistics for the class of 1973 are biased, as the scale was constructed using these individuals.

**Statistically significant biserial correlations (Alf & Abrahams, 1971):

Class of 1971: $r_b \geq .132$, $p \leq .01$.

Class of 1972: $r_b \geq .132$, $p \leq .01$.

statistics for the other cross-validation sample are somewhat better. Specifically, a biserial validity of .37 and a percent overlap of 75 were obtained for the class of 1972.

Expectancy charts illustrating the effectiveness of the Major Group II Scale for the classes of 1971 and 1972 are presented in Figures 3 and 4, respectively. Selection of those persons scoring in the top 60% of the distribution would result in approximately 46% Group II majors for the class of 1971, and 58% for the class of 1972. The corresponding base rates are 40% and 48%, respectively.

Major Group III Scale

Results of statistical analyses of the data on the Group III Scale are presented in Table 4. Cross-validation of this scale on the class of 1971 yielded a biserial correlation of .72 and a percent overlap of 49. Analyses of the class of 1972 data showed a cross-validity of .74 and a percent overlap of 45.

Figures 5 and 6 show expectancy charts for the Major Group III Scale for the two classes. For the class of 1971, persons scoring in the top quintile of the scale selected a Group III major over thirty times as frequently as did the group scoring in the lowest quintile, as shown in Figure 5.

The expectancy chart for the class of 1972, shown in Figure 6, closely resembles the chart for the 1971 class. Selection of those 60% scoring highest on the new Major Group III Scale would result in approximately 39% Group III majors for the class of 1971, and about 42% for the class of 1972. The base rate for both classes is 26%.

DISCUSSION AND CONCLUSIONS

Other Academy Classes

The authors consider the results reported herein as preliminary. The evidence obtained to date is necessary, but certainly not sufficient to justify operational use of the new scales. The relationship between scores on the scales and choice of major will be examined for the most recent graduating class (1974). In addition, members of classes currently enrolled at the Academy will be scored on the new scales and their scores will be evaluated in terms of their current major field.

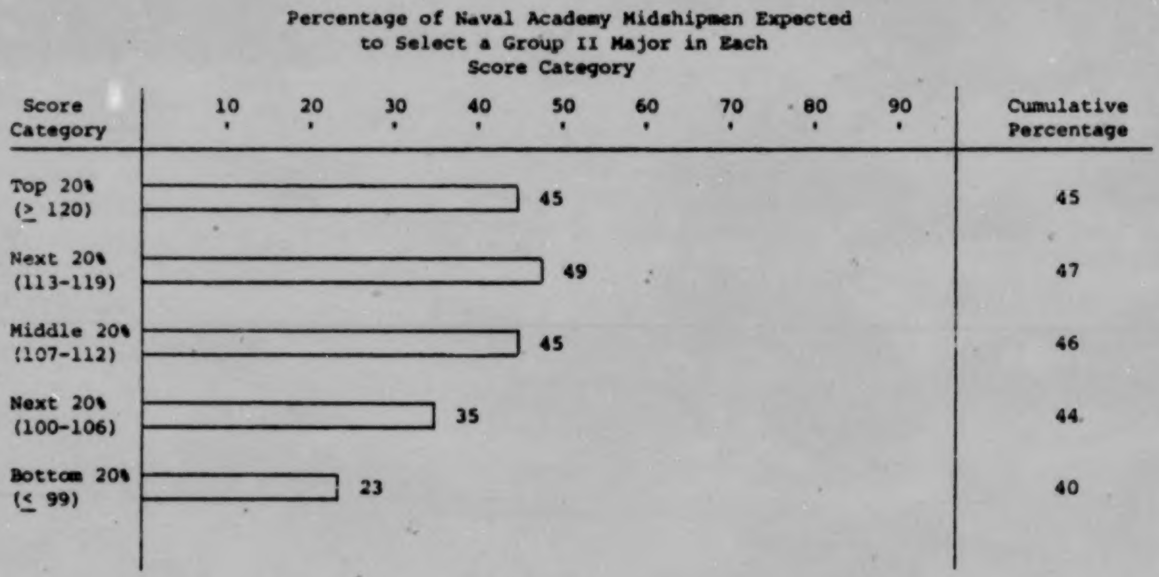


Fig. 3. Expectancy chart showing the relationship between the Major Group II Scale and choice of major for the Naval Academy class of 1971.

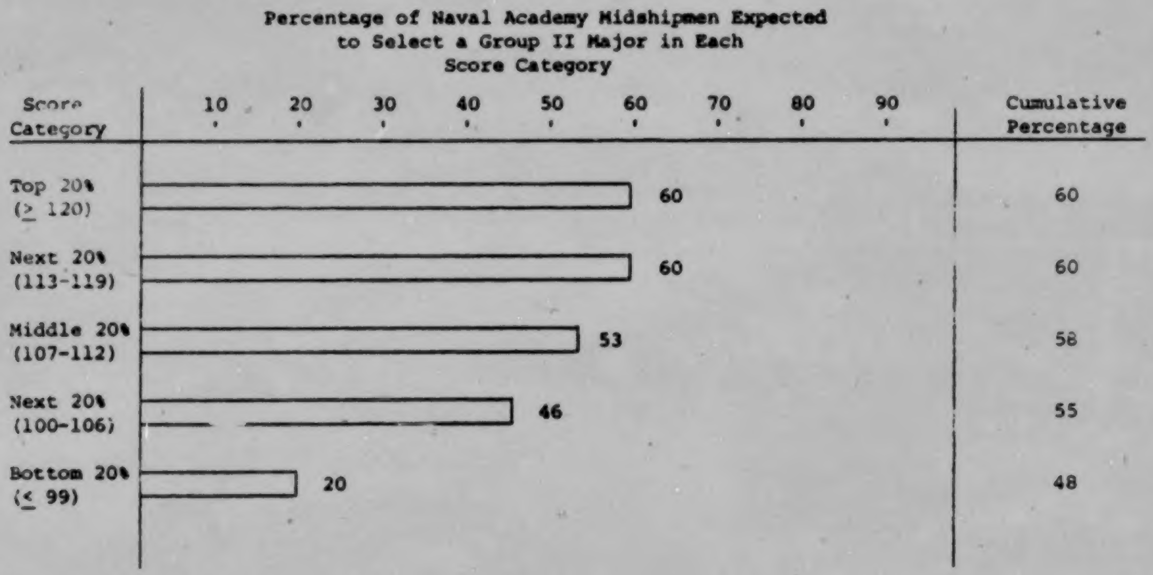


Fig. 4. Expectancy chart showing the relationship between the Major Group II Scale and choice of major for the Naval Academy class of 1972.

TABLE 4

Means, Standard Deviations, Validities, and Percent Overlaps for the Major Group III Scale on Naval Academy Classes of 1971, 1972, and 1973[a]

Class	Criterion Group	Major Group	N	Mean	Standard Deviation	Validity (r_b)	Percent Overlap
1971	High	III	171	106.47	17.61	.72**	49
	Low	I+II	489	83.01	16.06		
	Total	I+II+III	660	89.08	19.42		
1972	High	III	170	106.90	16.36	.74**	45
	Low	I+II	476	81.89	16.88		
	Total	I+II+III	646	88.48	20.04		
1973	High	III	178	110.24	18.96	.77[a]	45
	Low	I+II	485	84.09	15.75		
	Total	I+II+III	663	91.11	20.30		

Note.

[a] The statistics for the class of 1973 are biased, as the scale was constructed using these individuals.

**Statistically significant biserial correlations (Alf & Abrahams, 1971):

Class of 1971: $r_b \geq .141$, $p \leq .01$.

Class of 1972: $r_b \geq .142$, $p \leq .01$.

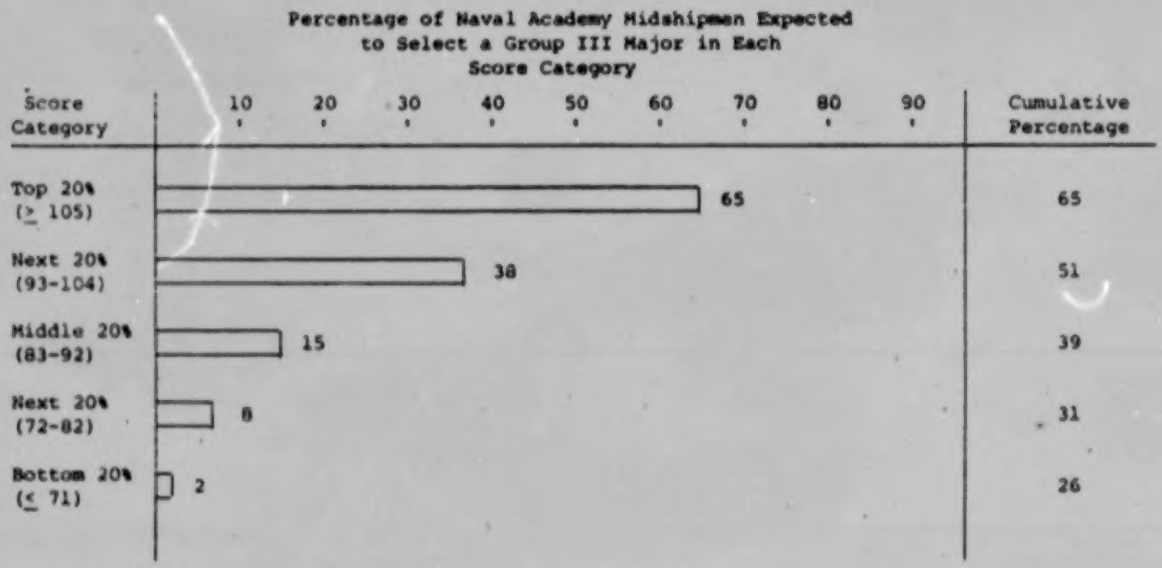


Fig. 5. Expectancy chart showing the relationship between the Major Group III Scale and choice of major for the Naval Academy class of 1971.

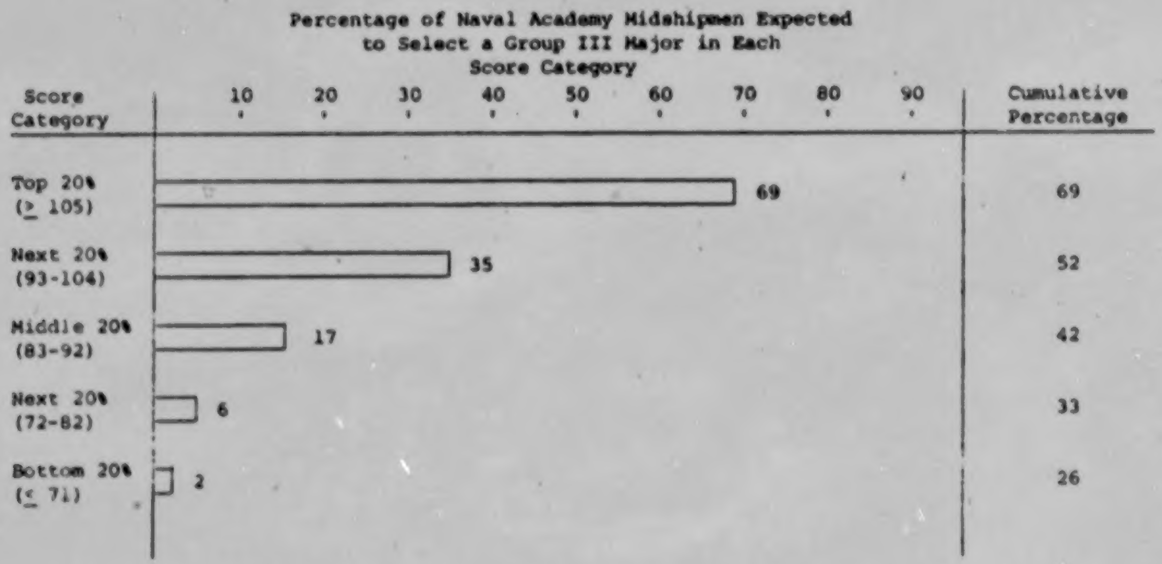


Fig. 6. Expectancy chart showing the relationship between the Major Group III Scale and choice of major for the Naval Academy class of 1972.

Test-Taking Set

For the classes of 1971, 1972, and 1973, the SVIB was administered to Academy midshipmen during the first week of plebe summer. Since the SVIB is currently part of the Academy application procedure, it is administered to applicants prior to entry into the Academy. Consequently, there is a need to determine if the difference in the two conditions (applicant vs. selectee) influences scores on the new scales. However, based upon previous experience, that is not likely to be a problem (Neumann & Abrahams, 1974).

The SVIB was administered to all applicants for the class of 1978. Then, during the first week of plebe summer, the inventory was administered to all selectees. As a result, pre-selection and post-selection SVIBs are available for approximately 1,400 persons. These SVIBs will be scored on the three scales to obtain some evidence on the influence of the "selection set" on applicants' scores.

Test-Retest Reliability

The SVIBs from the two administrations for the class of 1978 also will provide a basis for evaluating the test-retest reliability of the three new scales. The time interval between these two SVIB administrations was between six and eight months.

In addition, test-retest SVIB data are available for a longer time interval. The SVIB was administered to approximately 300 NROTC officers as applicants and then four years later as seniors.

Preliminary Evaluation

In conclusion, the preliminary results obtained for the new empirically-derived SVIB scales appear quite promising. The Major Group I Scale demonstrated high effectiveness. The Major Group II Scale was moderately effective. The Major Group III Scale was shown to be very highly effective. On the whole, the new SVIB scales performed well in both cross-validations.

REFERENCES

Abrahams, N. M., & Neumann, I. The validation of the Strong Vocational Interest Blank for predicting Naval Academy disenrollment and military aptitude. San Diego: U. S. Naval Personnel & Training Research Laboratory, February 1973. (Technical Bulletin STB 73-3)

Abrahams, N. M., & Neumann, I. Identification of Naval Academy applicants with engineering and science interests. Proceedings: Psychology in the Air Force Symposium. Colorado: U. S. Air Force Academy, April 1974.

Alf, E., & Abrahams, N. A significance test for biserial r. Educational and Psychological Measurement, 1971, 31, 637-640.

Campbell, D. P. Manual for the Strong Vocational Interest Blank. Stanford: Stanford University Press, 1966.

Campbell, D. P. Handbook for the Strong Vocational Interest Blank. Stanford: Stanford University Press, 1971.

Neumann, I., & Abrahams, N. M. Identification of Naval Academy applicants with engineering and science interests. San Diego: Navy Personnel Research & Development Center, October 1974. (Technical Report 75-7)

Tilton, J. W. The measurement of overlapping. Journal of Educational Psychology, 1937, 28, 656-662.

U. S. Naval Academy, The Majors Program. Annapolis: Academic Dean Notice 1531 of 1 July 1974.

INTEREST MEASUREMENT FOR ARMY CLASSIFICATION*

M.A. Fischl, Ph.D.

U.S. Army Research Institute for the Behavioral and Social Sciences

THE VIEWS EXPRESSED IN THIS PAPER/ABSTRACT ARE THOSE OF THE AUTHOR(S) AND DO NOT NECESSARILY REFLECT THE VIEWS OF THE UNITED STATES ARMY OR OF THE DEPARTMENT OF DEFENSE.

INTEREST MEASUREMENT FOR ARMY CLASSIFICATION

M. A. Fischl, Ph.D.
U.S. Army Research Institute for the Behavioral and Social Sciences

The purpose of this paper is to explain the philosophy of interest measurement underlying the tests of the Army Classification Battery, and to describe the way the interest measures are used in effecting enlisted classification.

The current Army Classification Battery, ACB-73, has been in operational use since May 1973 and was described thoroughly at MTA last year.[1] In brief, it consists of 12 separate subtests which yield 15 scores, combinations of which comprise the 10 aptitude area composites that the Army uses for enlisted classification and assignment.

Four of the fifteen scores are obtained from noncognitive vocational interest scales, and three more from cognitive tests of information about various trades. The philosophy of interest measurement implemented in both the noncognitive and the cognitive measures is to infer interest in a vocational activity from a demonstration of previous actual participation in the activity. In the noncognitive set a minimum number of items deal with what the examinee ''would like to do,'' and the maximum number of items deal with what the person has done or does like to do. Exhibit 1 illustrates the noncognitive items.

Where the cognitive items are concerned, the demonstation of previous actual participation in a vocational activity comes from information that the individual has retained about the activity. The tests of automotive information, electrical/electronic information, and construction trade information cover content that most people are exposed to in the normal course of growing up. Knowledge about any of these content areas is taken as indication that the individual was sufficiently interested as to pursue the activity, participate in it, and hence retain information not usually acquired and retained by people with different interests. As Exhibit 2 shows, the items of these tests are really quite easy for anyone whose interests have taken him in the direction of the particular trade, yet are virtually unanswerable by the person with no relevant experience. The point is that, although seemingly achievement or proficiency measures, they would really be too easy to discriminate well among people with any appreciable degree of experience in the line of work, but are about right to discriminate the person whose interests got him started in the activity from the person who has not participated.

---

[1] Maier, Milton H. Enlisted Selection and Classification Testing in the U.S. Army. Proceedings, 15th Annual Conference of the Military Testing Association, San Antonio, Texas, 1973.

Assignment to training in a particular job is made on the basis of aptitude area scores; that is, composites of from 3 to 5 of the 15 scales yielded by the battery. Nine aptitude area scores are used for classification purposes, a tenth is used to establish qualification for supplemental testing, such as the Officer Candidate Test. These aptitude area composites were established through multiple regression methods for predicting performance in entry level job training, later checked against on-the-job criteria. The noncognitive interest scales contribute to six of the composites; the cognitive interest subtests contribute to these six plus a seventh composite.

It should be noted explicitly that performance is the criterion that these measures are used to predict, and against that criterion the interest measures sufficiently enhanced the aptitude area multiple *R*'s as to warrant their inclusion along with the battery's measures of general and specific aptitudes. Exhibit 3 presents the zero order *r*'s for each of the interest scales, while Exhibit 4 presents the multiple correlation coefficients of the seven composites which contain the interest measures.[2]

In summary, the Army Classification Battery obtains seven vocational interest measures, three cognitive and four noncognitive. Interest in a vocational area is inferred from prior participation in the area and from knowledge obtained as a result of that participation. Scores on the interest measures are integrated with scores on the battery's ability measures, to yield composite scores from which differential vocational success is predicted and classification decisions made.

---

[2] All correlation coefficients in both exhibits utilize performance in entry level job training as the criterion variable.

EXHIBIT 1

ILLUSTRATIVE NONCOGNITIVE INTEREST ITEMS[1]

I have ‘‘gone backpacking’’ more than two or three times.

I am a pretty good carpenter.

I have balanced a checkbook.

I repair small electrical things around the house.

---

[1] Not actual items.

EXHIBIT 2

ILLUSTRATIVE COGNITIVE INTEREST ITEMS[1]

In an automobile the word ''points'' is associated with which system?

A) brakes
B) ignition
C) cooling
D) fuel

Usually the moving part of a direct-current motor is the

A) field
B) armature
C) brushes
D) frame

Crawler-type tractors are usually powered by

A) electricity
B) kerosene
C) gasoline
D) diesel fuel

---

[1] Not actual items.

EXHIBIT 3

MEAN VALIDITY COEFFICIENTS OF INTEREST MEASURES FOR PREDICTING TRAINING SUCCESS IN THE RELEVANT MILITARY OCCUPATIONAL SPECIALTIES

Interest Measure	Validity Coefficient
Noncognitive	
Ground Combat	.28
Mechanical Mainteance	.34
Electricity/Electronics	.44
Fine Detail/Attentiveness	.32
Cognitive	
Automotive Information	.46
Electrical/Electronics Information	.52
Trade Information (Construction)	.54

EXHIBIT 4

VALIDITY COEFFICIENTS, FOR PREDICTING TRAINING SUCCESS IN THE RELEVANT MILITARY OCCUPATIONAL SPECIALTIES, OF APTITUDE AREA COMPOSITES CONTAINING INTEREST MEASURES

Composite	Validity Coefficient
Ground Combat	.53
Field Artillery	.61
Electronics Repair	.73
Operators & Food	.45
Surveillance & Communications	.69
Mechanical Maintenance	.74
Clerical/Administrative	.68

SECTION III

FLIGHT SIMULATORS--
THEIR DEVELOPMENT AND VALUE IN TRAINING

SIMULATOR MOTION AS A FACTOR IN FLIGHT SIMULATOR TRAINING EFFECTIVENESS

Robert S. Jacobs
University of Illinois
Aviation Research Laboratory

The exceptional range of fidelity (and cost) of flight training simulators in use would suggest that there is an insufficient level of understanding of the elemental contributions of various cue dimensions to the overall effectiveness of the training device. This lack of complete understanding has usually resulted in selection and procurement of the highest fidelity simulation that the state of the art and budget can provide in an attempt to maximize effectiveness. Related research suggests that this objective may not have been realized in every case — particularly in a cost effectiveness sense. The one dimension of environmental fidelity which is probably most cost sensitive is that of motion cue structure. It is essential, then, to determine the relationship between motion cue fidelity and resultant training effectiveness in order to optimize the cost efficiency of the training system. An experiment in progress is an initial attempt to systematically develop this relationship for basic flight training applications.

The techniques by which a novice is educated to the skills required for the operation of an aircraft have been refined considerably over the history of flight training. Grahame-White and Harper (1916) describe the process as practiced in England in that time as consisting of a good deal of discussion on the ground between student and experienced aviator prior to the student's attempt to master the vehicle on his own. It is suggested that three of four flights were usually sufficient to prepare the student for his certification flight, which consisted of a series of passes around a closed figure-eight shaped course about two pylons "at least 200 yards apart." Having successfully demonstrated sufficient skill in guiding the aircraft around the course in the presence of an official witness from the British Aeronautical Society (the regulatory agency empowered at that time to certificate applicants for flying licenses), the student was awarded a pilot's license which permitted him, among other things, to instruct others in the secrets of the art of aviating.

Of course, as the vehicles themselves became more complex, the increased demands on the operators required a more comprehensive tutelage, certification process, and the imposition of standards with regard to the experience of and methods employed by flight instructors. The requirements of two great wars for the rapid training of a large number of skilled aviators prompted the pursuit of greater productive efficiency from the flight training process through the development of improved methods and training aids. A natural avenue of investigation in an era when the utilization of each aircraft diverted to flight training activities meant one less aircraft available for combat use was the development of training tools which could shorten the time required to achieve proficiency in the aircraft by substitution of practice in a suitable alternative environment. Efforts toward this end culminated in the late 1930s with Edwin Link's invention of the first widely used flight trainer.

The early flight trainers were little more than mechanical contraptions which grossly responded akin to aircraft to control input from the student pilot. As flight instruments permitting reliable all weather flight came into being, these appeared in the trainers to aid in the development of instrument interpretive skills. Advances in servo-mechanical engineering and electronics brought improvements in the technical ability to replicate aircraft dynamic response. At the end of World War II, according to Smode, Hall and Meyer (1966), simulators and other synthetic devices had come into wide use for initial skill training of student pilots, transition training, refresher training, skill maintenance, and training for specific types of missions. Adams (1961) points out that this employment of a flight training simulator can be rationalized on a number of grounds: economics favor the use of the relatively* inexpensive-to-operate simulator rather than the parent aircraft; the simulator may be used for teaching skills which are too complex, expensive, or risky to practice in the aircraft; the simulator affords a measure of control and the ability to isolate and practice particular segments of the overall flight task, whereas the aircraft does not; the simulator operates independent of prevailing weather conditions; and finally the simulator provides an opportunity to practice the operation of single place aircraft under supervision which would be impossible in the aircraft itself.

Transfer of Training

The measure of the worth of a flight training simulator in a specific training curriculum is established quantitatively by transfer effectiveness. The issue in deciding whether the training system should include a training simulator is the determination of the extent to which time spent in the simulator would reduce the need for training time in the aircraft to reach a stated level of proficiency. If simulator operating cost is less expensive than that of the aircraft by a ratio greater than the transfer ratio, a saving of overall cost may be achieved by employment of the simulator in the training system as evaluated. Transfer is often quantitatively expressed as follows:

$$T = \frac{C - E}{C + E} \times 100\%$$

where

C = trials to criterion in transfer task of control group
E = trials to criterion in transfer task of experimental group

---

* in the case of highly complex simulators, the operating costs although lower than those of the parent aircraft are none-the-less significant.

If the experimental group takes more time on the criterion task (practice to given standards of performances in the aircraft) than the control group, the numerator is negative, and an inhibiting, or negative, transfer of training is said to have occurred. If the two groups take exactly the same time to reach criterion, the ratio takes on the value of zero, and no transfer is considered to have taken place. On the other hand, if the practice of the experimental group prior to the criterion task does have a beneficial effect in terms of shortening the time to criterion performance, a positive transfer between 0 and 100% will have occurred.

There have been several researches of the transfer of training value of flight training simulators, the almost universal conclusions of which are that there is positive transfer of training associated with practice in a simulator. Many of the often cited examples were conducted at the University of Illinois' Aviation Psychology Laboratory during the late 1940s and early 1950s. For example, Williams and Flexman (1949) demonstrated that non-pilots could be trained to perform a series of maneuvers using a Link trainer and airplane in an alternating practice sequency with 28% fewer trials and 22% fewer errors than a group trained entirely in the aircraft. In a follow-on effort with expanded scope, Flexman, Matheny, and Brown (1950) demonstrated that a simulator used in a properly constructed curriculum could reduce flight hour requirements to achieve private pilot proficiency by more than 50%. More recently, Povenmire and Roscoe (1969) investigated the time savings impact of the use of the Singer Link GAT-1 trainer on the basic flight training curriculum here at the Institute of Aviation. They confirmed that the use of the simulator for practice has positive transfer to a Piper Cherokee 140B aircraft in this application, resulting in a savings of time in the in-flight practice segment of the course. Subsequently (1973), they investigated the relative transfer effect associated with a number of different practice ratios of GAT-1 and aircraft and found that there is a diminishing returns effect associated with time spent in the simulator past a certain point. They developed a measure which reflects this effect, incremental transfer effectiveness - the time saved in flight by each additional hour spent in the simulator. It was demonstrated that if the relative costs of training in the simulator and aircraft are known, it is possible to design a curriculum which minimizes total cost by selection of a most cost effective period of practice in the GAT-1. Other studies have examined the respective proficiency levels of groups trained in combined aircraft/ground-trainer and aircraft-only curricula (Mahler and Bennett, 1950; Wilcoxon, Davy, and Webster, 1954; etc.) with findings that the simulator trained groups were as or more proficient than aircraft trained groups with equivalent or less aircraft time.

It will be stipulated then, in view of this evidence, that the simulator does constitute a viable basic training aid, and that this is a widespread belief is evidenced by the fact that such devices are in common use throughout the basic flight education field. The beneficial transfer effects demonstrated in the cited studies apparently are sufficiently valuable to justify the outlay of the funds for procurement of training simulators by even the smaller of the flight training institutions, although the competitive pressures of this industry no doubt also had some influence here as well. These transfer effects are not, however, uniform across the entire spectrum of skill categories which together comprise the requirements for pilot certification. Ornstein, Nichols and Flexman (1954) analyzed the transfer effects of the use of a simulator in conjunction with aircraft as compared with the aircraft alone on a maneuver by maneuver basis to find that the simulator was most effective for certain classes of procedure loaded contact flight exercises, and least effective on maneuvers which the simulator did not well reproduce. They discuss the relationship between the transfer effectiveness of the simulator and the fidelity of simulator reproduction of the aircraft procedural and environmental cue structure, suggesting that by extending the range and fidelity of the simulation, transfer is maximized. This would logically follow by analogy from the Osgood Transfer Surface (Osgood, 1949) concept which hypothesized heightened transfer as a result of similarity between the practice and criterion tasks in the general transfer situation as opposed to the flight training context.

Fidelity and Transfer

This question of the relationship between the fidelity or realism of the simulator as compared with its aircraft counterpart as a determinant of transfer of training value became more important to the designers and purchasers of simulator systems as the state of technology advanced to make possible high fidelity simulation but only at a very great increase in simulator cost. Investigators began to examine the relationship between transfer value and realism along a number of simulator design dimensions. A basic and related investigative issue is the question of whether a simulator is really required, or whether an environment conducive to mental and physical practice of procedures without feedback might do as well. Dougherty, Houston, and Nicklas (1957) examined the training effectiveness of four alternative ground training devices including an operational flight trainer, a procedures trainer (without dynamic instruments), a photographic mockup, and the procedures trainer with an auxiliary side tracking task. These represent four levels of physical fidelity to the object aircraft. Each group was assigned one of these training aids for a program of procedures and maneuvers. Each then transferred to the aircraft and was compared in the airborne training program with a fifth group trained only in the aircraft.

All trainer groups performed significantly better on the first air trial than did the aircraft only group, but by the third air trial, the differences were not statistically reliable. The groups trained in the simulator and procedures trainer showed the highest degree of transfer, but again by the third air trial, no differences could be observed as a function of training method. The conclusion reached of importance here was that in addition to demonstrating that a ground trained group with either simulator or procedures trainer experience could be competitive with an entirely air trained group, the transfer of training could not be shown to be sensitive to the degree of physical and task fidelity in the trainer. Matheny, Williams, Dougherty, and Hasler (1953) had already demonstrated that faithful reproduction of flight control feel was not an essential characteristic of the flight trainer for reasonably high transfer to occur, and such studies in other training areas as that by Briggs, Fitts, and Bahrick (1956) demonstrated that control feel was not closely related to transfer of training for a generalized tracking task tend to contradict Osgood's notions (1949) about this similarity being critical for efficient transfer to take place.

Motion Cues and Simulator Fidelity

The accuracy of reproduction of the cue structure of an aircraft in flight as provided in flight training simulators has been greatly improved since the early attempts at synthetically generating these cues. Instrument indications and control pressures can be virtually identical to those of the object aircraft if sufficient resources are allocated to the computation of aircraft dynamic response and if sufficiently accurate models are employed. Real time, computer generated visual systems have enabled the simulator to provide a highly realistic analog of external visual cues, provided the customer can pay for these cues as well. Motion systems have also been developed which do an excellent job of generating acceleration and rotational cues to the simulator pilot which closely approximate those which may be measured in a real aircraft -- except for those cues of linear and rotational acceleration associated with rolling into, out of, and the maintenance of turns.

This cues fidelity problem will be easily understood by consideration of the physics of an aircraft in turning flight. When turning, the aircraft, in addition to being acted upon by the normal forces of flight; lift, drag, thrust, and gravity or weight, comes under the influence of a fifth acceleration force -- centripetal acceleration. The pilot within the aircraft is acted upon by this force as well. His vestibular system is incapable of differentiating between the component accelerations together composing the total motion cue environment, however, and he feels only the vector sum of all of the accelerations acting upon him. Thus in the aircraft, in a steady state turn, gravity and centripetal accelerations are perceptually combined to produce a resultant vector, which in a properly coordinated turn acts perpendicular to the plane of the wings.

The sensation for the pilot is one of increased weight alone, not one of turning or sliding across his seat. In a properly coordinated turn, the magnitude of this resultant vector, and hence his perception of weightiness is parabolically related to the angle of bank and consequently the rate of turn. In the simulator, however, only the force of gravity acts on the pilot unless the simulator cockpit is able to generate centripetal forces as, for example, in the case of a cockpit on the arm of a centrifuge. This will not be the case in most instances; the centripetal component of the acceleration environment will be missing. Where a centrifuge is used to generate the cues, the short radius of turn, as compared with the turn radius of an aircraft, produces other anomalies which detract from the realism of the motion cue set as well. The simulator thus will not reproduce accurately turning cues representative of an aircraft.

An approximation commonly used is the so called washout motion approach, in which the roll accelerations of the simulator cab are analogous to the roll accelerations of the aircraft, with exponential decay back to earth level at such a slow rate that a sustained turn is represented to the pilot by a level cockpit although he is not aware of the transition back from the banked condition. This is a compromise, however, and even though the direction of the resultant vector during a turn is appropriate with the washout motion, the magnitude is less than it realistically should be.

Even with this limitation, motion cues in the simulator do become a part of the total informational environment available to the simulator pilot. The processing of this information and the way that it is used to assist in the control of the vehicle may differ, however, from that adopted in the aircraft. If we consider the generalized vehicle control problem, we have a classical man-in-the-loop control system. The man is presented with various informational cues from his environment through his various senses. These cues are received, transformed, combined, processed and then used as imputs to a control model; we might term it an operative model; in order to determine what the state of the control process is, and what must be done to make the system conform to the desired state.

Motion cues have been hypothesized as playing two parts in the operative model of an aircraft. They may certainly be used as quantitative inputs to the control model to assist in the real time computation of appropriate control movements. They also may serve in an entirely independent fashion as a generalized alerting cue which signals the need for control action but is not used in the determination of what control action might be required. Flexman (1966) discusses a study which suggests that for the experienced pilot, the latter was the case. Student pilots trained to make instrument takeoffs in a motionless simulator performed with much greater success than did a group of experienced instructor pilots who it seems had grown to depend on motion cues as a signal that the aircraft attitude was changing. Even though these same instructors know full well that a continuous mechanical scan of the instrument panel is necessary to maintain control of an aircraft by instrument reference, and that in a real airplane vestibular cues are not reliable information sources in the absence of external visual confirmation of vestibular impressions, they apparently had come to depend upon vestibular cues as a signal to scan the instrument panel to discover the source and nature of the attitude change underway. Students trained in the motionless simulator had become much more dependent upon the mechanical scan, and did not so rely on vestibular alerting cues. They were, therefore, able to maintain control in the absence of them.

There are a few studies which compare pilot performance in motion equipped and motionless simulators and both of these with performance in an aircraft, (Douvillier et al., 1959; Rolfe, et al., 1966; Fedderson, 1961). In general it has been found that performance in a simulator with motion more closely approximates performance in an aircraft than does that observed in the fixed base simulator. Transfer of training comparisons, however, fail to support the notion that motion is an important contributor to increased transfer value. In the Fedderson study for example, although the motion trained students showed greater transfer to an aircraft on the initial aircraft trial, reliable differences had disappeared by the sixth transfer trial. Fedderson concludes that the use of a motion simulator was difficult to justify in view of this. Smode et al. (1966) review a series of

NASA studies examining the requirement for motion in simulators where the task was aircraft control for experienced pilots, and in this extra-training context, they summarize the findings as justifying motion only in the case of simulation of an aircraft where the motion cues present an essential stability contribution (for example, in the case of anticipatory cues), or where realistic motion is an inhibitory factor such as might be the case in vehicles with unusual dynamics or high vibration levels. Caro and Isley (1966) evaluted a tethered helicopter as a trainer for basic helicopter students and found that while those students who received the training performed better up to the first solo flight, all differences disappeared after this point and the conclusion was reached that the value of motion simulation of high fidelity is of limited value for long-term transfer purposes.

In each of the foregoing examples, performance differences were observed between groups of students trained on high fidelity motion devices and those trained on motionless simulators which initially suggest that the motion-trained students learn to achieve higher performance more quickly. However, an important distinction needs to be made; the function of the motion cues as learning vs. action feedback in the control task. Holding (1965) discusses the role of feedback in the learning process. He reviews a large number of studies in proceeding to the conclusion that the intrinsic, concurrent, and immediate nature of such performance cues as motion feedback of control inputs is more promotive of high performance than of rapid learning. It would seem that motion-trained students are presented with a practice environment which is richly endowed with aiding cues facilitating performance of the practice task at very high levels relative to that provided the no-motion student.

Koonce (1974) conducted a study which supports this contention. Koonce was attempting to validate the use of a ground-based simulator as a proficiency evaluation environment by comparing performance in the simulator with performance in an aircraft. He found that those pilots presented with the more difficult task of performing the check ride maneuvers in the motionless simulator performed less well in the simulator than those tested with motion, but that this difference was radically reversed in the aircraft (see Figure 1). The difficulty of the task for the motionless simulator group apparently served to strengthen skills required for superior aircraft performance to the detriment of simulator performance. The criterion in this instance is aircraft performance level however. It seems clear that if any skill increase was taking place, as the negative slope in Figure 1 would suggest, the difficult task, without motion, was a better learning environment than the easier motion case.

Let us consider the notion of an operative model once again. If we accept the previously discussed general concept of the man-machine control loop as being a useful model for discussion, we may state that it is the task of the student in basic training in any skill to derive empirically an operative model by associative experimentation with input information and trial and error control outputs. Of course this process is speeded by the presence of an instructor, but the student must experiment

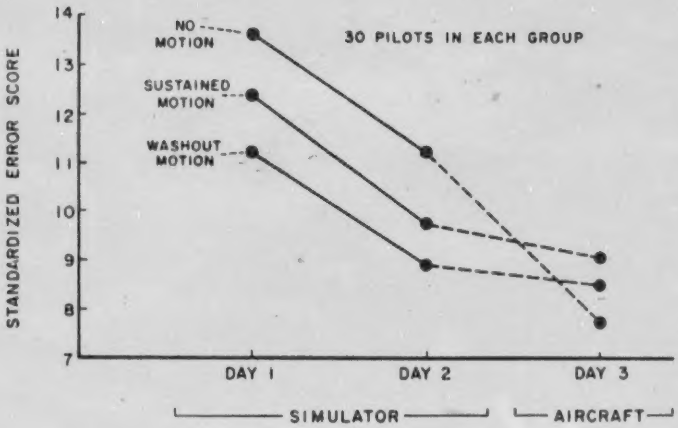


Figure 1. Performance in an instrument flight skill evaluation under three conditions of simulator motion (Koonce, 1974).

with the quantitative aspects of control response even if the qualitative aspects are provided by instruction. He must learn how hard to push the stick forward to break a stall, how soon to anticipate the flare point by pulling the stick back and with how much force, even if he knows qualitatively beforehand that these are appropriate actions. It is only reasonable to assume that in the formation of an operative model, the student will utilize all of the relevant and correlated informational cues available to him to make the task as easy as possible.

Thus if a student is trained in a simulator which provides a cue structure that is more abundant than that provided by the aircraft, the student will include these extra cues in his operative model, and when removed, the lack of these cues will require adaptation to maintain performance levels. What will happen, however, if the student believes that all the cues required for the operative model are present when in fact they are not, or are in fact present but false. Faulty control must result.

This is precisely the situation which will arise when a student pilot is trained in a simulator which supplies highly correlated and accurate motion cues, comes to depend upon them, and then attempts to fly an aircraft under instrument referenced conditions which often produce misleading vestibular indications. Would it not have been better for this particular student to have been trained under a condition of no motion cues at all, so that his operative model makes no use of these misleading cues? Perhaps, but what if he is incapable of ignoring them entirely? There is no point at which we can say that learning has ceased and performance has begun. Thus it is possible that the operative model will be continuously modified to take advantage of cues available. Is this process of adding to the complexity of the model more inhibiting than revising it to exclude previously available terms? This is a dilemma which has yet to be explored fully.

What Is Being Done?

A survey of the simulation devices in common use for basic flight training will confirm the impression that the manufacturers and buyers of this class of device have very little idea of the component contributions to transfer effectiveness of the various dimensions of cue fidelity. With respect to motion cue fidelity, it is possible to procure systems which provide either no, poor, or quite good but imperfect roll motion simulation at prices which correspond to the fidelity offered. If roll motion fidelity is an important determinant of transfer, the purchasers of the no-motion systems may not be getting an appropriate return on their investment. If not, as some of the earlier cited research seems to indicate, money spent on motion systems is wasted. The question of how much roll motion fidelity is required to produce cost effective transfer has not been satisfactorily answered.

The fidelity of roll motion is not a single dimension of measurement, but rather a multidimensional variable itself. In addition to the very basic question of whether there need be any roll motion at all, the character of the motion cues may be just as important an issue. In part, the answer will rest upon the use to which the cue is put in the operative model. If the roll motion cue serves only an alerting function, then any perceivable intensity of stimulation in any direction should serve as well for a naive interpreter. If, however, the magnitude, temporal character, and direction of the cue are also important, motion cues which are then uncorrelated with attitude state in any of these dimensions could not be used to produce useful corrective control inputs by way of the operative model process. The experienced flight instructor will caution the student to ignore motion cue direction and magnitude because in the aircraft these can be misleading, even though as Flexman (1966) found, the instructor apparently will not or cannot ignore them himself.

It becomes important then in addition to examining the impact of the motion cue structure on transfer, to distinguish whether these cues are being used by the naive trainee as alerting or informational signals. At the Aviation Research Laboratory of the University of Illinois' Institute of Aviation, a study is in progress intended to disclose the function and significance of motion cues in the simulator. We are running a classical transfer experiment involving four groups of flight naive subjects. One group serves as a control and is being trained to perform a series of basic instrument referenced maneuvers in a Piper PA-28R-200 Arrow single engine light aircraft. The other three groups receive preliminary instruction in a Singer Link GAT-2 training simulator prior to beginning the instructional sequence in the aircraft. These groups receive their simulator training under three motion conditions; none, high fidelity washout motion, and directionally uncorrelated random-roll washout motion respectively.

If it is found that the third and fourth groups have equivalent transfer performance, it will be apparent that the directionally random washout motion which can serve only as a general alerting signal is as effective in facilitating aircraft control as is correlated, highly realistic motion. If, however, the higher fidelity motion provides reliably greater transfer, it will be apparent that the direction and magnitude of the cues have a purpose in the operative model as well. If the no-motion group transfers as well as the two motion groups, the value of motion cues of either variety as an adjunct to transfer will be questionable. The aircraft only group will serve as a standard for judgment of the value of simulator practice as abstracted from the motion cue structure.

The Training Task

If differences between these groups do in fact occur, they will be amplified by operating at the early, steep portion of the learning curve. Naive subjects are being taught the most basic elements leading to the first major milestone in pilot training, the solo flight. Exercises representative of this early training are taught on a trial basis, with total trials to reach criterion performance and total errors to reach criterion performance serving as the measures for analysis. These elements include procedures and techniques for climbs, glides, straight and level flight, turns, and basic instrument referenced flight.

A recording and scoring system used by Williams and Flexman (1949) is used. The technique calls for the instructor to judge whether each trial on a maneuver met stated standards of performance along critical dimensions. if a predetermined number of successful attempts are made by the student, the instruction advances to the next phase. Failure to meet criteria requires repeat of the trial.

Subjects

Subjects are volunteers who have no previous piloting experience. Since aptitude for flight training is a measure which exhibits considerable individual variation, two means were used to balance experimental groups. First, a sufficiently large sample will be made to minimize the effect of these differences. Twelve students will be assigned to each of the four groups, making a total of 48 subjects for the experiment. Second, each of the potential subjects has been tested with a battery of flight training aptitude measures, and groups will be balanced across these. The aptitude tests will include a divided-attention time-sharing performance task involving simultaneous tracking and number processing, a mental arithmetic test, and the Guilford-Zimmerman Pt. V Spatial Aptitude test. Seven measures resulting from these tests are used to assign students to groups. A discriminant analysis of the subject population scores has shown that the groups are properly matched with respect to each of the seven measures.

Instructional Technique

Several flight instructors are involved in the experimental training program. To minimize the effect of instructor variation, material is presented to the students largely in the form of audio and video tape cassettes with a minimum of real-time interaction between student and instructor. The instructors function principally as maneuver scorers and safety pilots.

Skill Areas Sampled

Fitts (1962) considered the skill content of such perceptual-motor processes as aircraft control and concluded that there are three principle constituent activities; spatial-temporal patterning, continuous interaction of response processes with input and feedback processes, and learning. While learning will pervade the other processes in this program, measures and maneuvers have been selected which will sample the development of each of the first two activities. Maneuvers included in the curriculum will draw on (1) basic motor control and the development of a "feel" for the airplane as indicated by skill with the primary controls and the trim tab, (2) perceptual skills as indicated by instrument referenced flight performance, and (3) supporting mental and imaginational skills as revealed by instrument referenced orientation and navigation.

Analysis

An analysis of variance based upon total trials to criterion performance will be used to compare performance by treatment group to examine the transfer effects of the various motion conditions. In addition, a past hoc examination of the predictive validity of each of the seven balancing measures against total trials to criterion across skill areas will be made to examine the effectiveness of the group matching measures. Correlations of matching measure scores with individual maneuver scores will further amplify the predictive value and limitations of these measures.

Results

At the conclusion of the effort, it will be possible to state with greater certainty how the gross characteristics of flight trainer roll motion cues relate to the transfer of training effectiveness of the device when applied in a basic training program. Although this is only a first step in the total process of delimiting the total contribution of specific motion cue characteristics to the training value of the simulator, the course for future research will be more clear. If it can be shown that the transfer of training value of the simulator is in fact relatively insensitive to the presence or absence of these cues, or to the fidelity of their reproduction where present, the implications for flight training will be very significant. With quantitative data available as to the degree of transfer which may be expected from simulators with varying characteristics, the simulator designer or user may for the first time make the choice of how much fidelity to include in his device on a rational basis. The prevailing custom of letting the budget fix the level of realism in the simulator will be supplanted by systematic maximization of the transfer effectiveness within the constraint of cost. The research evidence to date suggests that this maximization when determined in a cost effectiveness framework may occur well below the cost levels associated with very high fidelity motion systems.

A very realistic possibility is that the characteristics desirable in the motion system of the simulator will be found to vary as a function of the training objectives it will be expected to meet. We have earlier seen that a given simulator will produce differential transfer across a maneuver set with differing skill components. It is anticipated that this study will reveal that the provision of high fidelity roll motion simulation may facilitate the learning of simple motor control and tracking to some extent, but if Koonce's (1974) findings are generalizable to the instrument referenced flight situation high fidelity may hinder learning as compared with no motion or directionally random motion conditions. With respect to the development of the imaginational components of piloting skill, it is difficult to predict whether motion characteristics will have any transfer implications beyond those associated with heightened realism and possibly resultant increased motivation of the student.

REFERENCES

Adams, J. A. Some considerations in the design and use of dynamic flight simulators. AFPTRC-TN-57-51. In Sinaiko, H. W., Selected Papers on Human Factors in the Design and Use of Control Systems. New York: Dover, 1961.

Briggs, G. E., Fitts, P. M., and Bahrick, H. P. Learning and performance in a complex tracking task as a function of visual noise. USAF: AFPTRC-TN-56-67, 1956.

Caro, P. W., Jr. and Isley, R. N. Helicopter trainer performance following synthetic flight training. Journal of American Helicopter Society, 1966, 11 (3), 38-44.

Dougherty, D. J., Houston, R. C., and Nicklas, D. R. Transfer of training in flight procedures from selected ground training devices to the aircraft. Savoy, Ill.: University of Illinois, Institute of Aviation, Aviation Research Laboratory, Technical Report NAVTRADEVCEN 71-16-16, September 1957.

Douvillier, J. G., Jr., Turner, H. L., McLean, J. D., and Heinle, D. R. Effects of flight simulator motion on pilot performance of tracking tasks. Washington, D. C.: National Aeronautics and Space Administration, Technical Note D-143, 1959.

Fedderson, W. E. Effect of simulator motion upon system and operator performance. Paper presented at the 7th Annual Army Human Engineering Conference, Ann Arbor, Michigan, October 3-6, 1961.

Flexman, R. E., Matheny, W. G., and Brown, E. L. Evaluation of the school Link and special methods of instruction in a ten-hour private pilot flight training program. University of Illinois Bulletin, 1950, 47 (80) (Aeronautics Bulletin No. 8).

Graham-White, C. and Harper, H. Learning to Fly - A Practical Manual for Beginners. New York: McMillan and Company, 1916.

Holding, D. H. Principles of training. New York: Pergamon Press, 1965.

Koonce, J. M. Effects of ground based aircraft simulator motion conditions upon prediction of pilot proficiency. Savoy, Ill.: University of Illinois at Urbana-Champaign, Institute of Aviation, Aviation Research Laboratory, Technical Report ARL-74-5/AFOSR-74-3, April 1974. [Also: Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1974.]

Mahler, W. R. and Bennett, G. K. Psychological studies of advanced naval air training; Evaluation of operational flight training. New York: The Psychological Corp., Technical Report SDC-999-1-1, September 1950. (AD 643 499)

Matheny, W. G., Williams, A. C., Jr., Dougherty, D. J., and Hasler, S. G. Effect of varying control forces in the P-1 trainer upon transfer of training to the T-6 aircraft. Savoy, Ill.: University of Illinois at Urbana-Champaign, Institute of Aviation, Aviation Psychology Laboratory, 1953.

Ornstein, G. N., Nichols, I. A., and Flexman, R. E. Evaluation of a contact flight simulator when used in an Air Force primary pilot training program: Part II. Effectiveness of training on component skills. Lackland Air Force Base, Tx.: AFPTRC-TR-54-110, December 1954. (AD 62373)

Osgood, C. E. The similarity paradox in human learning; A resolution. Psychology Review, 1949, 56, 132-143.

Povenmire, H. K. and Roscoe, S. N. An evaluation of ground based flight trainers in routine primary flight training. Human Factors, 1971, 13, 109-116.

Povenmire, H. K. and Roscoe, S. N. The incremental transfer effectiveness of a ground based general aviation trainer, Human Factors, 1973, 15, 537-545.

Rolfe, J. M., Hammerton-Fraser, A. M., Poulter, R. F., and Smith, E. M. B. Pilot response in flight and simulated flight, Ergonomics, 1970, 13 (6), 761-8.

Smode, A. F., Hall, E. R., and Meyer, D. E. Assessment of research relevant to pilot training. USA: USAMRL TR-66-196, Wright-Patterson AFB, Ohio, November 1966.

Wilcoxon, H. C., Davy, E., and Webster, J. C. Evaluation of the SNJ operational flight trainer, Port Washington, N.Y.: SDC TR-999-2-1, March 1954. (AD 086 988)

Williams, A. C., Jr. and Flexman, R. E. Evaluation of the school Link as an aid in primary flight instruction. University of Illinois Bulletin, 1949, 46 (71) (Aeronautics Bulletin No. 5).

DEVELOPMENT OF AN RPV OPERATOR TRAINER

Louis E. Ehrhardt

United States Naval Missile Center
Human Factors Engineering Branch
Point Mugu, California 93042

INTRODUCTION

The requirement for realistic targets for the test and evaluation of modern weapons systems, and for the training of the weapon system operators, has increased considerably from the simple targets provided for ship's gunnery practice during World War I. As Naval weapons systems grew in sophistication and complexity, the philosophy employed in target design moved away from floating bullseyes to vehicles which provided a realistic simulation of the threats the weapons systems were designed to counter.

In the present test, evaluation, and training communities, the simulated threats span the range from stationary sea and land targets to high speed boats and high performance aircraft. However, techniques devised decades ago for the operation of such targets are still in common use today. Most operating equipment and techniques now employed are the result of piecemeal assembly and chance evolution and few, if any, have considered human engineering principles in design and usage.

Because of the similarity of air vehicle targets and their associated control problems to the emerging weapons systems designated as RPVs (Remotely Piloted Vehicles), it was felt that human engineering problems attendant in current target operations could provide useful insight into potential problems in RPV operations and provide a rational basis for further exploration of RPV control and display configurations.

Under the sponsorship of the Office of Naval Research and the Naval Air Systems Command, the Naval Missile Center undertook a study to determine which control and display characteristics optimized operator performance on the RPV control stick. This study centered on the RCO (Remote Control Operator) training device in use at the Naval Missile Center for the training of Naval drone operators.

This paper discusses the development of this training device and the implications of student performance for future RPV control station design.

Trainer Development

Historically, the only training a novice RCO received was on-the-job training. The novice would observe an experienced operator for many flights while the operator explained the meaning of the displays, the functions of the flight and special purpose controls, and the flight characteristics and idiosyncrasies of the particular airframe. During non-critical portions of an operation, the novice would be allowed to practice controlling the drone and learn the "feel" of the control inputs. After a gradual transition from observer to hands-on controller, a novice would be pronounced a qualified RCO and would be included in the normal operations schedule.

Unfortunately, this mideval apprenticeship method of training did not adequately prepare the new operator for special or emergency situations. Drone problems such as thrust misalignment, autopilot malfunctions, or unsymmetrical aerodynamic surface alignment, and sudden failure, such as loss of telemetry or ordnance damage, generally resulted in inappropriate action on the part of new operators and consequent drone loss. As the complexity and cost of the targets increased, the Naval Air Systems Command decided to develop a formal training program and tasked the Naval Missile Center to develop the curriculum and conduct the classes.

In response to this task, the Naval Missile Center designed and built an RCO training device and structured a curriculum which not only provided the students with operational practice, but also the theory of operation of the drone and its subsystems. Since most Naval RCOs have not had previous flight training, it was determined that a period of formal classroom instruction devoted to principles of flight, flight terminology, and aerodynamics as well as the drone subsystems would be beneficial. This has proven to be the case both during the training period and the subsequent operational period. Having received approximately twenty hours of classroom instruction in basic aerodynamics and the airframe and subsystem characteristics of the drone, the students are exposed to approximately eighteen hours on the drone control simulator.

It was the simulation concept which made the entire program possible and cost effective. Simulation has long been used, quite successfully, for flight training, particularly instrument flight training. However, flight simulators have traditionally been criticized for lacking several real-world flight cues, primarily lack of proper "g" force and lack of realistic "out-the-window" visual presentations. Several attempts have been made to overcome these difficulties, but the controversy continues. I mention this point because the simulation of an RCO crew station and the attendant tasks does not face these problems. When performing his task, the RCO sits in a closed room using a standard control box and synthetic flight displays. To simulate this operational environment, all that needs to be provided is the control box and the appropriate displays and, as those of you who have ever been in an operations room know, a rather large amount of distracting background noise.

The simulation was mechanized in the Naval Missile Center's main computer simulation laboratory - thus providing both the necessary equipment and background noise. The heart of the simulation is an EAI 231R analog computer. The analog approach was chosen as the simplest and cheapest alternative capable of meeting the realtime operator interaction requirement and the interface requirement for the controls and displays used in the operator console mockup. The equations which govern the kinematics, dynamics, autopilot, and control systems of the drone were mechanized on the analog mainframe. The analog computer also contained the interface and buffer circuitry necessary to accept inputs from the control box and to provide driving signals to the displays.

A digital computer, an XDS SIGMA 2, was used to generate the necessary aerodynamic functions and the autopilot gain schedules that are dependent on altitude and dynamic pressure. The digital and analog computer were interfaced such that the digital program continually samples the appropriate analog variables, calculates the functional values, and outputs the values to the analog computer. This hybrid (Analog + Digital) computer system constitutes the analytical model portion of the overall simulation.

The next major element of the simulation is the operator's control station mockup. This consists of a console which contains the drone control box and metered drone status information such as velocity, angle of attack, pitch and roll angles, etc., and a large X-Y plotter for displaying relative ground track and altitude. The console contains two groups of displays. The first group is used for classes devoted to the BQM-34A drone, commonly called the Firebee I, and the second group is used for classes devoted to the BQM-34E supersonic drones, commonly called the Firebee II. Similarly, the console contains the control boxes particular to each of these drones. A third control box for the MQM-74A drone is integrated with the simulation, but not with the console. This will be explained shortly.

The primary display for virtually all target operations is the X-Y plotter. As stated earlier, this shows the drone's position relative to the fixed earth - usually a preplotted North-East flight path. It is the operator's primary task to keep the target on this path at a prescribed altitude, displayed by the second arm on the plotting board, and at a given airspeed. Once the student has mastered the basic control concepts, a process that requires about three hours of simulator time, this task is relatively simple. It is at this point that the third element of the simulation assumes importance.

The third element is the instructor's problem box. It is this component of the training device that makes the simulation approach far superior than the old apprenticeship method. The problem box allows the instructor to interface with the simulation so that he can cause both minor and major simulated problems which an RCO might experience during actual drone operations. Examples of such problems are: roll, pitch, and yaw biases that range from just noticable to very serious; axial thrust misalignment; and selective or catastrophic loss of telemetry.

By having to cope with these problems on the simulator, the student learns to recognize the onset of the problem and the appropriate corrective action. The failure of the old instruction method to teach these two elements was responsible for the loss of several vehicles annually.

Grading Students - The Problem

The purpose of training is to impart to the student some knowledge or skill that he does not have. However, all people do not learn all lessons equally well and instructors usually attempt to rank the proficiency of their students along some scale they feel is relevant to the particular learning task. In specialized skills, such as the remote control of airframes, the important criteria tends to be a yes - no situation: did the student learn the task? or did the student not learn the task? The subjective opinion of the instructor, based on his observation of the student's performance, supplies the outside world with the answer.

The two instructors conducting the RCO classes have a combined experience of over 200 students, but have "failed" only three or four (to the best of their recollection). However, Operations Conductors freely admit that some RCOs are very poor controllers and should not be included on the schedule until they have received more training. The instructors realize that some of their "graduates" perform better than others, but they feel that all are capable of the job.

Personally, I am reminded of the engineer that designed the bridge that collapsed of its own weight before construction was completed. He had his "yes" answer (a Bachelor's degree), but further investigation showed that it was granted with a C- average.

I do not mean to imply that the instructors or the controllers are not trying to do the best possible job. The point is, objective performance criteria do not exist against which a student's performance may be measured and a decision made to either provide more training, drop him from the curriculum, or graduate him.

RPV Research - A Backdoor Solution

In 1974, the Human Factor Engineering Branch was tasked by the

Office of Naval Research (Psychology Programs, Code 455) to investigate control-display relationships for the RPV control task. The objective of the research was to determine which displays and which control types gave rise to the best control performance by drone controllers. The motivating factor for this research was a curious statistic that came from the operational targets community.

During calendar year 1973, the Navy conducted 372 BQM-34A and BQM-34E operations. Of this number, 17 or 4.6% were recorded as unsatisfactory because the target deviated too far from the prescribed flight path and the deviation could not be ascribed to equipment failure. During this same time period, 649 MQM-74A operations were conducted. However, with almost twice as many operations, there were no recorded instances of unsatisfactory presentations due to excessive deviation of the target from the prescribed flight path. The natural question is: Why the difference?

Let us return to the control boxes mentioned earlier. Both the BQM-34A and E drones employ the same control philosophy. The operator has a four position discrete switch which allows him to command a climb or dive and a left turn or a right turn and a separate switch which allows him to command straight and level flight. However, once the operator has initiated a maneuver, the rate at which the maneuver proceeds is entirely under the control of the autopilot. Since the autopilot uses a set of schedules dependent upon airspeed, altitude, and thrust, the operator is seldom sure of the exact behavior of the drone as manifested by the track on the plotting board. Thus, the controller must constantly switch between an anticipatory mode and a "wait-and-see" mode of behavior. He must anticipate the time at which to make the all-or-nothing decision and then must wait until the plotting board displays the results of that decision. Since, from the standpoint of the operator's control input, there is no difference between, say, a 1.2 "g" turn and a 2.5 "g" turn, turns must be made using a series of turn and straight and level commands. This requires considerable practice and a fine sense of judgment on the part of the operator.

In constrast to the Firebee control philosophy, the MQM-74A control system employs a proportional concept. The operator has two knob controls, one for pitch angle and one for roll angle. The use of this control requires a better understanding of aerodynamic principles on the part of the operator (i.e., combinations of pitch and roll angle to achieve a turn) than does the discrete approach. However, positional control allows the operator to make fine adjustments in the maneuver being performed as the situation requires. It was suspected that this control difference was responsible for the difference in the unsatisfactory presentation statistics. This suspicion was reinforced by the fact that in the

Fleet, the MQM-74A is flown with no instrumentation other than an altitude readout and a much smaller plotting board which leads to less display resolution which normally leads to greater inaccuracies.

A series of experiments were devised to determine the most appropriate control philosophy for the remote piloting task. Four displays were chosen as being representative of those proposed for RPV missions. The displays were:

1. Fixed Map - same as the current target drone display.
2. Moving Map - same map as above only with a fixed vehicle symbol and a moving background.
3. Synthetic Video - A computer generated symbolic representation of the three dimensional world.
4. Real World Video - A television presentation of the World as would be seen by a forward looking T. V. sensor.

For each of these displays, three types of rate control and three types of position control were to be evaluated by a common group of qualified RCOs. However, in order that comparisons between control-display configurations could be made, a performance criteria and measurement method had to be developed.

As stated above, the primary task of an RCO is to maintain the vehicle on a prescribed flight path. Therefore, the ability of the operators to accomplish this was taken as the performance criteria. The specific measures used were the integrated errors squared for the X, Y, and Z deviations. Since the experiment is still in progress, the results cannot be reported at this time. However, there are some important implications for assessing student performance.

Using the above measures in conjunction with an airspeed deviation measurement, it would be possible to provide a quantitative assessment of the student's ability to perform the routine control task. By testing the performance of currently qualified RCOs, a scale could be developed against which comparisons made and a more objective decision reached as to the suitability of the student for the job. Additional data, particularly learning rates, could also be established. While this data may not be necessary at this time, because of the small number of people in the RCO community, it will become important as RPVs find their way into our weapons inventory.

The quality of the RCOs we train in the coming years, the quality of our defensive forces, and the effectiveness of our weapons systems ultimately depends on the quality and objectivity of the data we use to make decisions.

PERFORMANCE EVALUATION

THE USE OF SCORING SYSTEMS IN ADAPTIVE TRAINING

By

Squadron Leader C. J. Hyatt, RAF and Captain O. H. DeBerg, USAF

INTRODUCTION

This paper is intended to describe work which we have done at the Crew Station Design Facility, in the field of scoring complex tracking tasks. It is intended to provide a down-to-earth approach to the development of a philosophy for scoring systems without getting deeply involved in mathematical approaches or analysis or results. The typical sort of tracking task we might look at in the facility is landing an aircraft in Instrument Flying Conditions, using various forms of approach aid, and with various task loadings.

The approach we use to this problem will obviously read across fairly readily to many other areas of study involving the use of motor skills. Our work to date has been entirely in the field of developmental studies rather than training, but since both these processes are amenable to treatment by an adaptive loop approach, they have many features in common, and in particular they both need a meaningful scoring system. Future projects which will require us to develop scoring systems specifically for adaptive training are the Simulator for Air to Air Combat, and the Advanced Simulator for Undergraduate Pilot-Training.

Figure 1 shows how the Crew Station Design Facility fits into the ASD organization. It is part of the Directorate of Crew and AGE Engineering, which is responsible for providing an advisory service to the individual systems program offices of ASD. You will see that we are in the same division as the simulator branch, and we also have simulators of our own, used almost exclusively for experimental work.

As the emphasis on simulators in training increases, and these simulators become more sophisticated, the need for sophistication in our scoring systems increases, and over the last year we have found it necessary to educate our 'customers'. This paper is a distillation of our studies.

GENERAL DISCUSSION OF PHILOSOPHY

I want to start by looking at some definitions; working definitions rather than academic ones. The first of these is what we mean by an adaptive process. Figure 2 shows how we have defined it for our purposes in very general terms.

Figure 3 shows how this process can be represented by a classic three block loop, and I think most workers in this field would accept this general concept. More specifically in the context of training, the three blocks can be more precisely defined as shown in the lower half of the figure. There already appears to be a fair consensus of opinion that the quality of the performance measure is "the make or break" feature of the system. Most people would agree that it is unfortunately also heavily subjective. To make the performance measure as good as possible let us first look at what we want from a scoring system. We have made a list (Figure 4) of qualities which appear to be important:

> The first and most important is that the scoring system should be directly related to the objectives of the adaptive process. To ensure this, we have to persuade the training or experimental director to define his objectives clearly.
>
> The remaining qualities are not in any particular order: If the objectives are multiple, each aspect must have some element of the scoring system directed toward it.
>
> Many parameters may be collected during a study, and the mass of data can be confusing. It is necessary to reduce this to manageable and comprehensible proportions.
>
> Although some subjective decisions must be made in the formulation of the scoring system, the user should not have to make any when applying it-ideally a computer should be able to handle it.
>
> Having achieved a numerical value for the score, it should be possible to relate the various ranges with an acceptability index, for example: Excellent, Good, Average, Poor, Unacceptable.
>
> For the greatest benefit to be obtained from the adaptive system, knowledge of results is a valuable aid, and hence it is desirable if possible for the scores to be generated in real time.

To fully apply these principles we feel that it is necessary to develop the performance measure block of the classic adaptive loop (Figure 5). The classic adaptive training loop calls for the application of the philosophy of training in the formulation of the adaptive logic. This means simply how the next step of training is governed by present performance. Because this is the only block labelled 'Logic' there is a tendency to try and use this step to insert logic related to the training objectives. We believe that the performance measure block should be developed into three separate stages as shown in the lower part of the figure, and that the value logic and adaptive logic should be kept entirely separate from each other. It is these three steps that we want to concentrate your attention on. In simple and direct terms these three steps can be relabelled as shown in Figure 6. Our aim is to develop the philosophy which should be applied to the generation of the value or scoring logic step in this process.

DEVELOPMENT OF SCORING LOGIC

Having defined the concept of 'scoring logic', it is easy to see that this logic is the primary link between the raw data which can be obtained by monitoring performance, and the score which is used to determine subsequent progress.

It is the sole point in the loop at which the recorded performance is weighed against the fundamental objectives of the process. As a result, the quality of the scoring logic is the key factor in determining whether or not the adaptive cycle is efficiently directed toward the aims of the process, be it an experimental study or a training program.

It is only too easy to skip over the question of objectives when defining the logic, and there is, perhaps, an even more insidious risk of using existing scoring logic 'because it worked well last time'. Scoring systems tend to look similar to each other, especially those used in any one particular field of endeavor, and subtle but vital differences can go un-noticed.

One way of minimizing this risk - which we believe is a worthwhile investment of time - is to carry out an objective evaluation of the true aims of every new scoring system we devise, and to develop a sound rationale for the scoring logic. Better still, this rationale should be formally written up and included as an integral part of our description of the scoring system.

Let us then look at the various ways in which the aims of the adaptive process we are considering can affect the way we go about scoring it. There are a number of questions we have to ask ourselves, and some of the major ones are outlined in Figure 7.

The first - and in our developmental studies the most fundamental - question, and yet oddly enough the one most frequently overlooked, is which part of the man-machine system are we looking at: the man, or the machine, or the interface between them.

In the training context the answer is simple - the man, that is to to say the trainee, is paramount. I would like to digress for a moment, though, and speculate on the wealth of data which has, at one time and another, been collected within adaptive training systems and which if it is still stored, could have potential value for the study of the training machines used, or of the way they display information. It seems probable that much of this data might be tapped simply by running it through new scoring systems with appropriate changes in their aims - assuming that the original scoring system was correctly aimed at training.

Much valuable information on the merits or shortcomings of various monitoring or operating consoles currently in use might be acquired in this way.

However, as I said, this is a digression, and in the training context it is the trainee we are trying to assess.

The next question we have to ask is what are we trying to find out about the trainee. The basic answer here is obvious: we want to know how well he performs. But to decide how to measure this performance, we need to ask a number of subsidiary questions.

For example, we have to decide which of the parameters available to us - which means those parameters we can measure without undue expense - are relevant to performance. And the performance we use as a yardstick here must itself be performance which is directly relevant to the training objectives.

A good example of the sort of decision we have to make in this area arises when scoring a tracking task such as an Instrument Landing Approach. Should we consider distance along track as a measure of performance? This distance is one way of looking at the trainee's control of his speed; not just his instantaneous speed, but the integral of his speed with respect to time from the start to the present. Thus if he makes an error in speed, to minimize his resulting penalty score he must not only correct the speed, but make a suitable compensatory adjustment to bring him back to his correct position along track. The argument against using this parameter for scoring is that as long as the trainee stays on the correct line through space as defined by the landing system, it does not matter when he gets to various points along it. However, this obviously depends on the scenario. If the object of the exercise is not only to fly along the correct line and land at the correct point, but also to land at a specific time to fit in with an existing traffic pattern then the accuracy of his position along track must be considered of some importance.

Another decision we have to make, which is also related to what we are trying to find out about the trainee's performance, is whether we should be concerned with his continuous operating or tracking ability, or only with his ability to reach a certain point by any means at his disposal. This quite clearly will determine whether we want a continuous scoring system or what we term 'gate' scoring - that is to say a measure of his ability to pass through a gate in space, or perhaps a series of gates.

I realize that you may be thinking that the points we are making here are overly simple and obvious - they are simple, and they should be obvious, but we think it is vitally important to emphasize that a cold-blooded analysis of this sort should be made, rather than the sort of approach which we know from experience often does occur, based on the principle of "Well, we usually score parameters A, B, and C - it looks as if they should be OK again this time."

Having, we hope, established which parameters we want to select from the raw data, we next have to think how the values of these parameters can best be used to give us a measure of performance. The logical approach is to compare these achieved values with some ideal. Most scoring systems adopt this approach, but once again a vital step which may be missed is to ensure that the ideal we set is truly relevant to our objectives. It is useless to set an ideal value of some parameter at 100 feet plus or minus zero, when we know that the 'Ace of the base' can only achieve 100 feet plus or minus ten feet, and plus or minus twenty feet is quite adequate for routine performance of the task. This is also a useful point at which to consider the limitations of our measured data; we can run into serious problems if we try to score to the nearest foot, when the equipment - and I include here both the instructor's monitoring equipment and the trainee's operating equipment - can only measure to the nearest five feet.

Even having established an appropriate ideal and a valid measure of divergence from it, we still need to consider what the implications of this divergence are.

First of all we must look at the relationship between size of divergence and importance - for example in a certain situation a five foot error could be acceptable, a ten foot error merely embarrassing, but a fifteen foot error fatal. Clearly this is not a linear relationship - at least, not in terms of human values, and our score should reflect this. In an extreme case of this sort of situation, any error less than ten feet might be totally acceptable, while anything in excess of ten feet would be totally unacceptable.

This is an example of what is generally known as 'time on target' scoring. In some contexts such as air to air combat with guided missiles it may be a perfectly adequate measure of performance, although even here it is more suited to competitive scoring than to training. But for a complex tracking task it contains too little data to help the instructor or trainee to identify areas of weakness and plan remedial training accordingly.

Another thing we have to take into account when assessing the divergence of the achieved performance from the ideal is the fact that some parameters may be much more critical than others, and if we want a scoring system which assigns equal penalties to equally unacceptable errors, we must weight the measured errors accordingly.

A FORMAL PROCESS FOR CREATING A SCORING SYSTEM

I now want to recapitulate on the various sorts of decision we have discussed, and in so doing I want to formalize and develop a process for creating an effective scoring system.

I want to show how, starting from the raw data available to us, we can adopt a systematic approach to mould it to our purposes, and ensure that our training objectives are met.

There are various processes we can apply to a mass of raw data, but we believe that six of these processes, shown in Figure 8, if applied in sequence will go far toward producing a sound system.

The first step is to select the parameters which are relevant to our training objectives.

Next we must look at what data is available to us on these parameters and edit it. By editing it, I mean deciding which values of it to use, and the choice here ranges from using it all, through using it at regular intervals, to using it only at specific points which we consider relevant.

We must then compare this edited data with what we believe to be the ideal values we are seeking to achieve by out training program. This comparison will give us what we have chosen to term 'error values'.

These error values must now have two processes applied to them. These are Modification and Weighting. The dividing line between them is not clear cut, and for this reason - to avoid lengthy discussion of what each comprises, I will not separate them. The operations which I include under these headings are:

Ensuring that the error values reflect the trainee's performance rather than any shortcomings in the training equipment.

Ensuring that the size of the error and the importance of this size is suitably reflected in the score.

Ensuring that the more critical parameters carry appropriately heavier scoring penalties.

Ensuring that any inter-relation between parameters is accounted for - this for example would include any weighting in respect of range if this were considered relevant.

The result of modifying and weighting the error values is to produce what we call 'scoring elements'.

Finally the scoring elements we have arrived at can be combined to give a single comprehensive score, or they may be combined in groups to give sub scores related to particular parts of the training objective, or particular capabilities of the trainee. Part of this combination process will be normalization of the score with respect to time or distance if this appears to be appropriate.

We believe that this systematic approach gives the best chance of achieving a useful and meaningful scoring system.

PRACTICAL EXAMPLES

If our systematic approach is applied to a variety of training programs, a variety of scoring systems will naturally result, but in all probability they will fall into four or five broadly defined groups; for example time on target systems, cumulative error systems, gate score systems, or combinations of these. Within any of these groups, one can postulate a generalized scoring system which by manipulation of a series of constants can be used for various slightly different training tasks. This concept lends itself very well to a computer based scoring system within an evolving organization - for example a pilot training school - where the training objectives remain fairly constant but the equipment used, and the associated operating procedures, will probably evolve steadily over the years.

To give you a good example of this, I will offer a brief outline of the type of scoring system we are currently using in our developmental studies to assess a pilot's ability to carry out landing approaches under instrument flying conditions. Remember of course that in these studies our aim is to assess the man-machine interface: This does not affect the process of scoring, but only the detailed application of the scoring logic.

We have concluded that the appropriate method to adopt in this instance is continuous scoring, with the score a function of size of error. Time on target scoring simply does not tell us enough about the things we need to know. However we do also take several specific gate scores at appropriate points en route, depending on the nature of the particular study. Our terms are defined in Figure 9, and a generalized formula for this continuous scoring is shown at Figure 10.

You will see that this formula allows for easy adaptation to suit changes in the relative importance of the different errors by adjusting the constants K, changes in the relationship between magnitude and importance of errors by adjusting the functions of the error values, changes in normalization philosophy by adjusting the function of time, and changes in weighting for range by adjusting the range function.

To illustrate how this type of formula has been applied, I will use the Microwave Landing System as an example. The object of this exercise is to ascertain whether or not certain tracks in airspace can be flown, on a time schedule, in a safe and efficient manner. Different routes are to be flown, and the object of the scoring system is to indicate the relative desirability of each route. Some typical routes are shown in Figure 11.

An MLS route differs from the standard ILS approach in several ways. First it is time dependent. The pilot must be in the right place at the right time. Also he must fly several different headings, some on a command basis, some by dead reckoning. Finally the descent rate is not constant, but depends upon which portion of the approach is being flown.

Scoring these approaches depends upon our value logic (as is true in any scoring system). This logic is based upon the objectives defined for the study. As initially set forth the purpose of MLS approaches is to control aircraft throughout a given airspace both with respect to time and spatial orientation relative to the prescribed path. This leads to the following logic:

(1) The aircraft must be equally controlled throughout the airspace.

(2) The aircraft must be in the "right place at the right time."

(3) For safety considerations the further the aircraft is from track the closer it approaches a critical situation.

(4) Some types of error are more critical than others (i.e. low on altitude is worse than high on altitude).

(5) Different tracks must be compared to one another on the same basis.

Apply these criteria, all possible parameters are analyzed and either discarded or modified and included in the score as considered appropriate. The formula shown in Figure 12 is the result. No range weighting is included because he must fly just as accurately at great distances as he does in close. Time of arrival considerations are met by using along track error which is time dependent. Hence the error in this direction is taken to be the distance of his actual position from where he should be. The safety consideration states that big errors can be critical, hence a square law is used. This penalizes large errors very heavily. The weighting constants were arrived at subjectively by discussion with qualified personnel as to the relative importance of different types of errors to be considered in a safe approach. Finally, in order to compare different tracks the entire score is normalized with respect to time. The resulting equation can be taken as a whole or by parts to examine the quality of the approach.

CONCLUSION

To conclude, we do not claim to have found all the answers, but we do feel we have gone a long way toward asking the right questions.

We have described a learning process that we have gone through, and which we suspect many people go through in the process of producing scoring systems. It has been a salutory experience for us to formulate our ideas into a systematic process. We hope that our presentation today will stimulate interest in the process and perhaps save others some of the time we have spent.

Remember: To produce an effective scoring system three things are essential. Good raw data must be collected, good value logic, or scoring logic, must be applied to it. The resulting components must be assembled in a practical manner. Figure 13 shows how our proposal sequence of operations generates the first three steps of the five step adaptive loop. The common thread running through the whole process is relevance - we must constantly ask ourselves if our scores relate to our aims.

We also hope that we have given you some food for thought, and that some of you will feel like contributing your own ideas in discussion now. Perhaps we may discover a few more of the answers to the questions we have posed. Thank you.

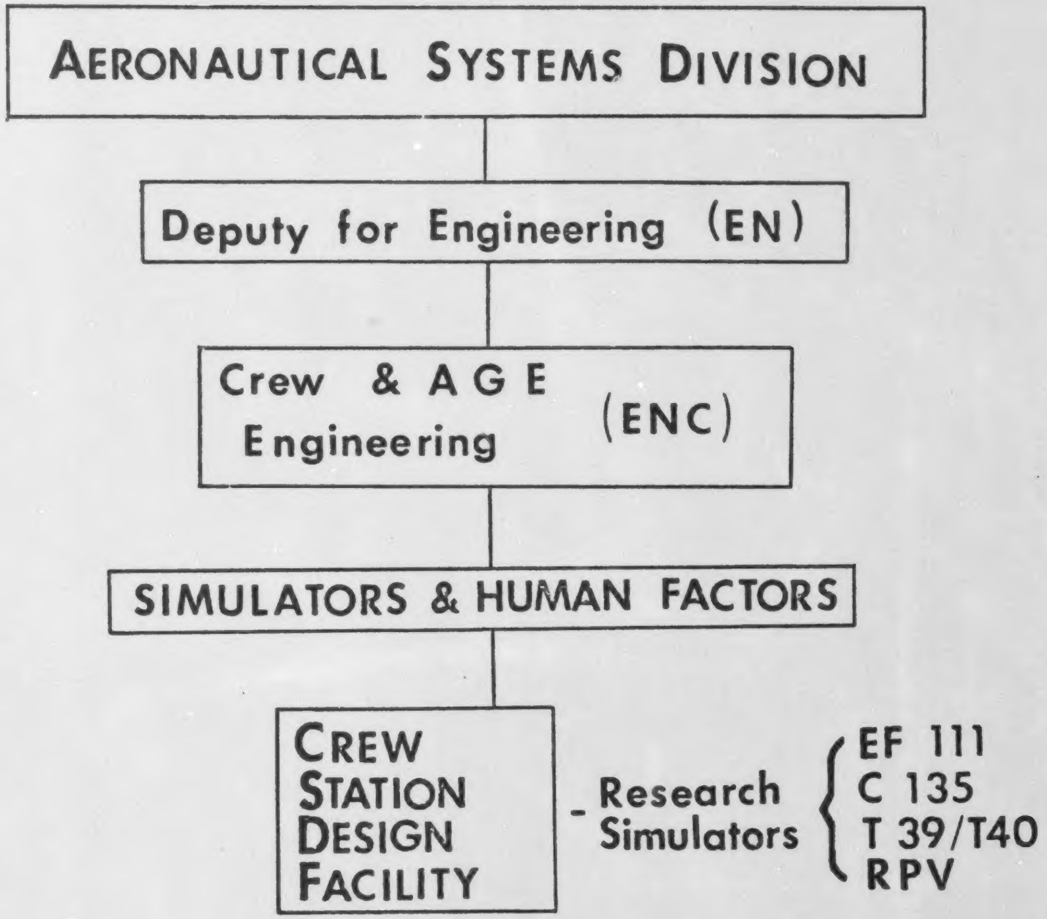

AERONAUTICAL SYSTEMS DIVISION
Deputy for Engineering (EN)
Crew & AGE Engineering (ENC)
SIMULATORS & HUMAN FACTORS
CREW STATION DESIGN FACILITY
- Research Simulators
EF 111
C 135
T 39/T40
RPV

FIG 1

THE ADAPTIVE PROCESS

A PROCESS IN WHICH THE OUTCOME OF ANY STEP DETERMINES THE NATURE OF THE NEXT STEP.

THE RELATIONSHIP BETWEEN ONE STEP AND THE NEXT IS GOVERNED BY THE APPLICATION OF A PREDETERMINED LOGIC.

FIG 2

THE ADAPTIVE LOOP

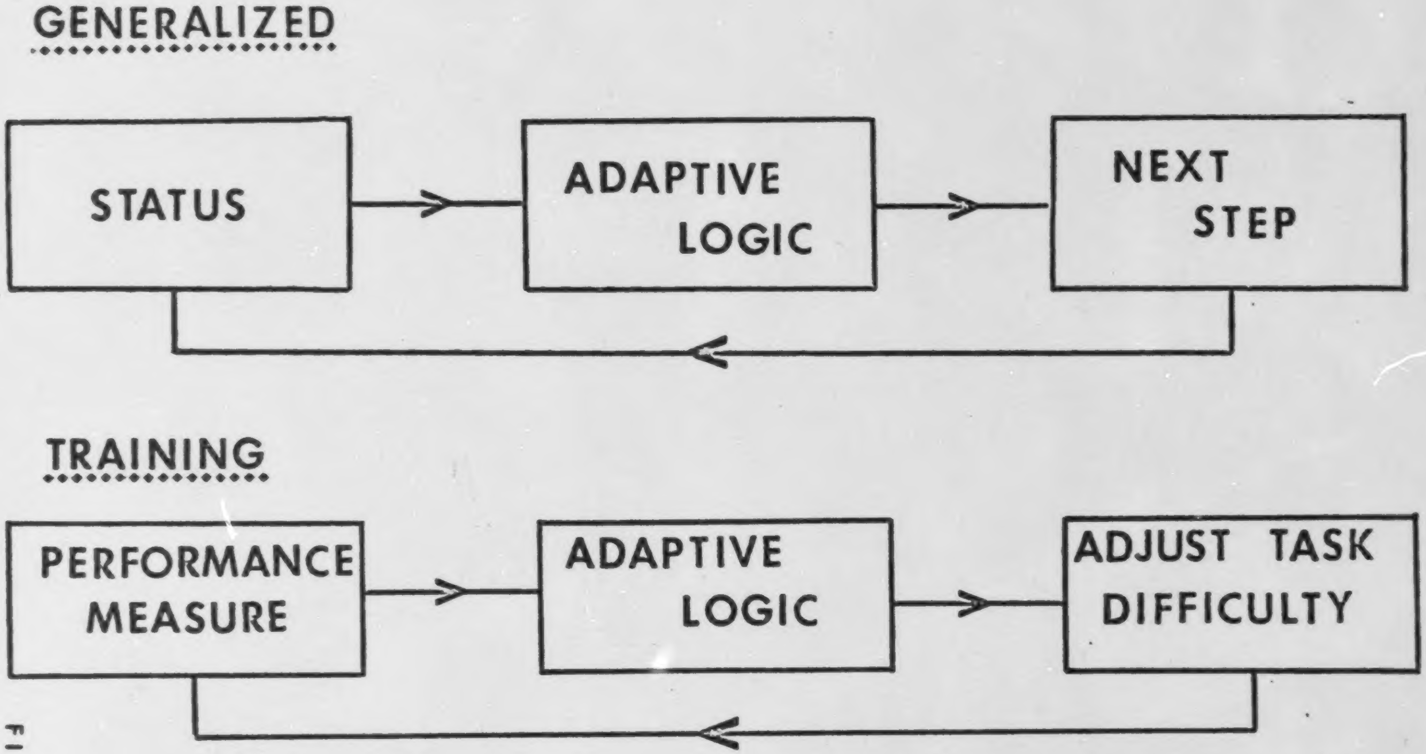

GENERALIZED
STATUS
ADAPTIVE LOGIC
NEXT STEP
TRAINING
PERFORMANCE MEASURE
ADAPTIVE LOGIC
ADJUST TASK DIFFICULTY

FIG 3

AN EFFECTIVE SCORING SYSTEM SHOULD:

★ BE DIRECTLY RELATED TO OBJECTIVES

★ DEAL WITH ALL ASPECTS OF OBJECTIVES

★ CONDENSE RAW DATA INTO SIMPLE MEASURES OF CRITICAL SYSTEMS ACTIONS

★ NOT BE DEPENDENT ON SUBJECTIVE MEASURES

★ RELATE TO AN ACCEPTABILITY INDEX

★ PRODUCE RESULTS QUICKLY, PREFERABLY IN REAL TIME

FIG 4

THE ADAPTIVE TRAINING LOOP

Development of Performance Measurement.

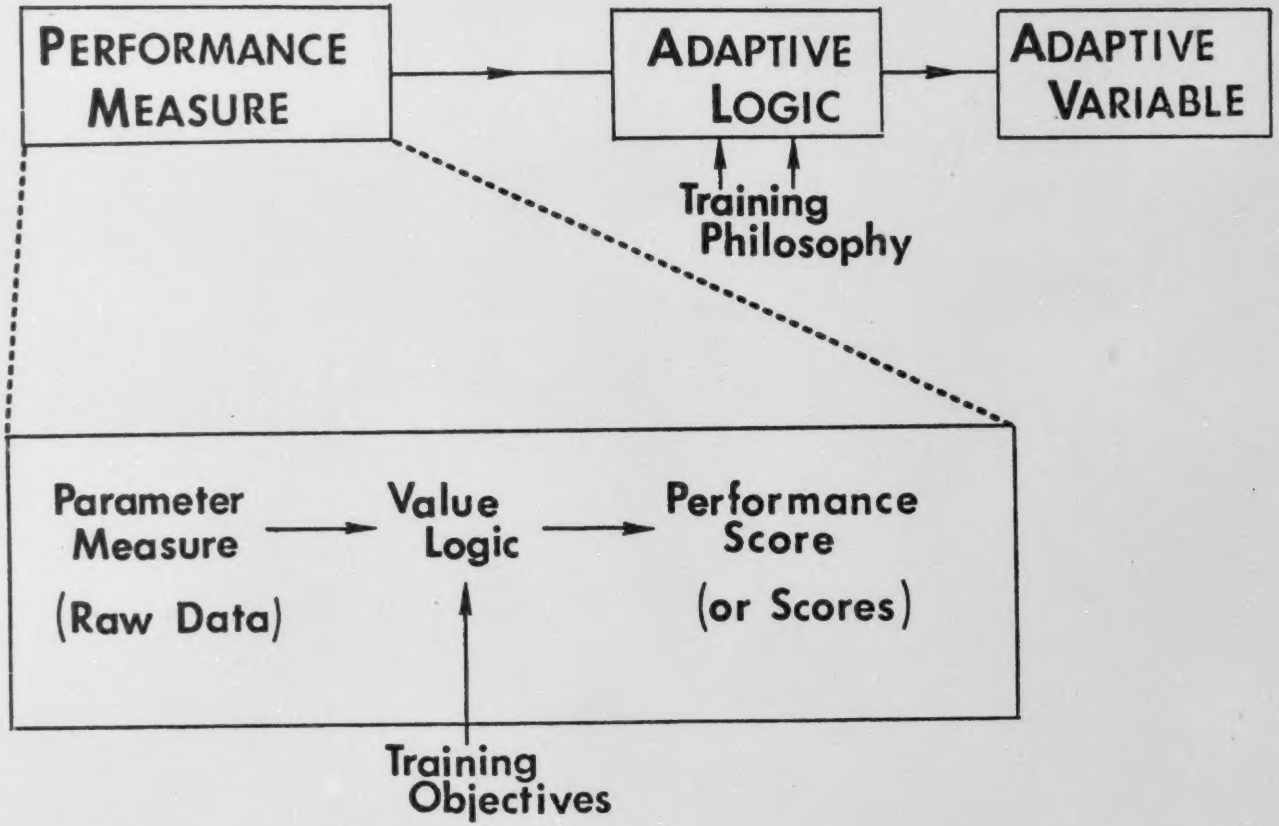

PERFORMANCE MEASURE
ADAPTIVE LOGIC
ADAPTIVE VARIABLE
Training Philosophy
Parameter Measure
(Raw Data)
Value Logic
Performance Score
(or Scores)
Training Objectives

FIG 5

ESSENTIAL ELEMENTS OF A SCORING SYSTEM

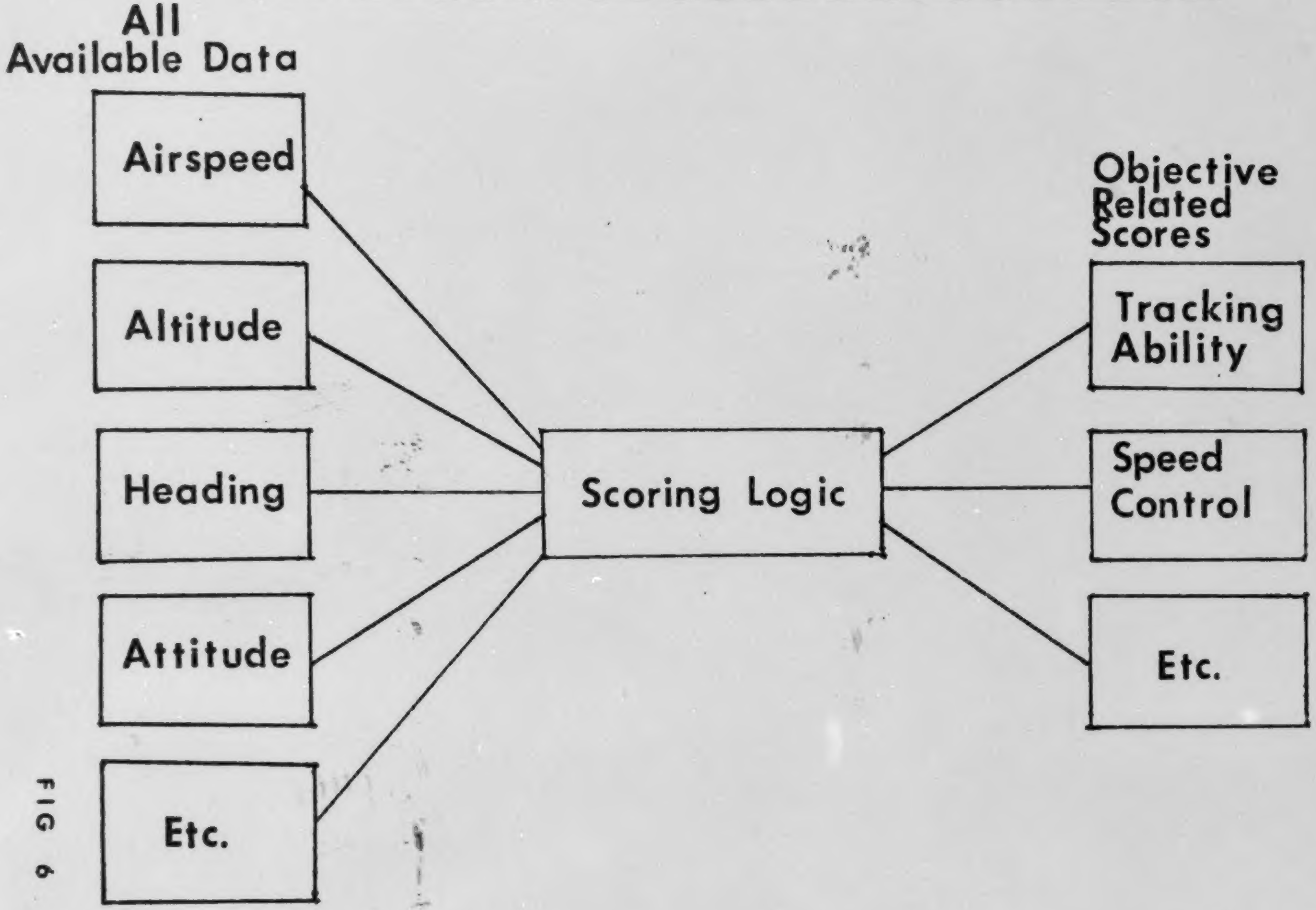

All
Available Data
Airspeed
Altitude
Heading
Attitude
Etc.
Scoring Logic
Objective
Related
Scores
Tracking
Ability
Speed
Control
Etc.

FIG 6

SCORING LOGIC MUST BE RELEVANT TO OUR AIM

Q. Are we looking at the man, the machine, or the interface?

A. In the training context - the man.

Q. What do we want to know about him?

A. How well he performs.

Specifically, Which parameters are affected by performance?

Which parts of the performance are relevant?

Q. How can we measure his performance?

A. Compare it with an ideal.

Specifically, How precise must the ideal be?

Are some apparent errors actually equipment limitation?

How are size and importance of error related?

Are some parameters more critical than others?

FIG 7

APPLICATION OF SCORING LOGIC TO RAW DATA

- SELECT
- EDIT
- COMPARE WITH IDEAL
- MODIFY
- WEIGHT
- COMBINE

FIG 8

SCORING SYSTEM DEFINITIONS

SUBSCRIPTS

x - (Along Track)
y - (Across Track)
z - (Vertical)
t - (Time)

TERMS

$E \longrightarrow$ Error Value
$F \longrightarrow$ Scoring Function
$R \longrightarrow$ Range
$\left.\begin{matrix} k \\ K \end{matrix}\right\} \longrightarrow$ Constants
$t \longrightarrow$ Time

Typically:

$$f(E) = E, E^n, \log E, E-K$$

$$f(R) = 1/R, 1, R$$

$$f(t) = t, 1, kt$$

FIG 9

The Scoring

Simplified,

$$S = \sum_{t=t_o}^{t_f} F_x + F_y + F_z \qquad \text{where} \quad F = \frac{kf(E)f(R)}{f(t)}$$

Extended,

$$S = \sum_{t=t_o}^{t_f} \frac{k_x f(E_x) f(R)_x + k_y f(E_y) f(R)_y + k_z f(E_z) f(R)_z}{f(t)}$$

FIG 10

TYPICAL MLS FLIGHT PATHS

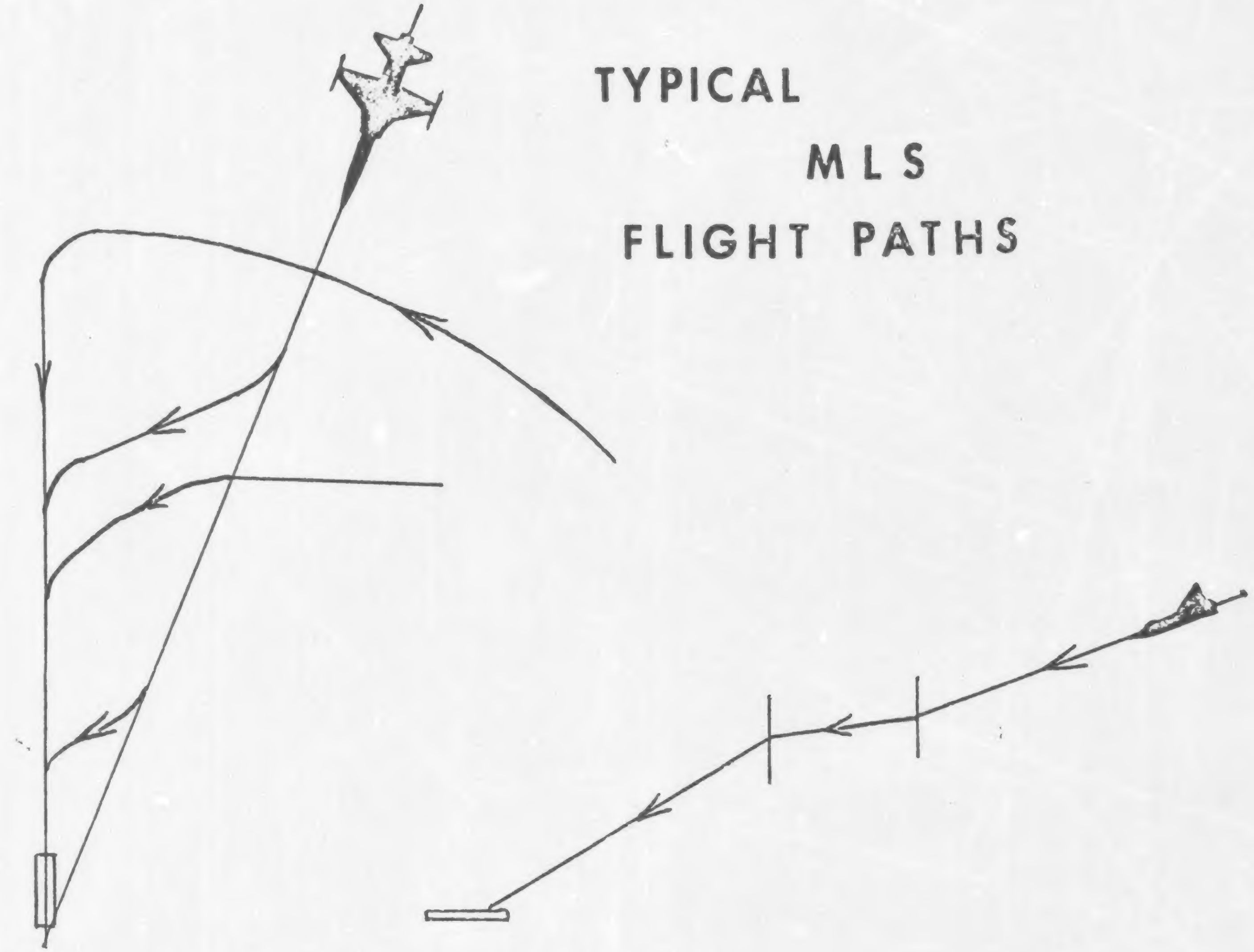


FIG 11

MLS Scoring

$$S_{MLS} = \sum_{t=t_o}^{t_f} \frac{\left(\frac{1}{32} E_x(t)\right)^2 + \left(E_y\right)^2 + \left(k_z E_z\right)^2}{t}$$

where

$$k_z = \begin{cases} 2 & \text{if} \quad E_z > 0 \\ 3 & \text{if} \quad E_z < 0 \end{cases}$$

FIG 12

THE COMPLETE PROCESS

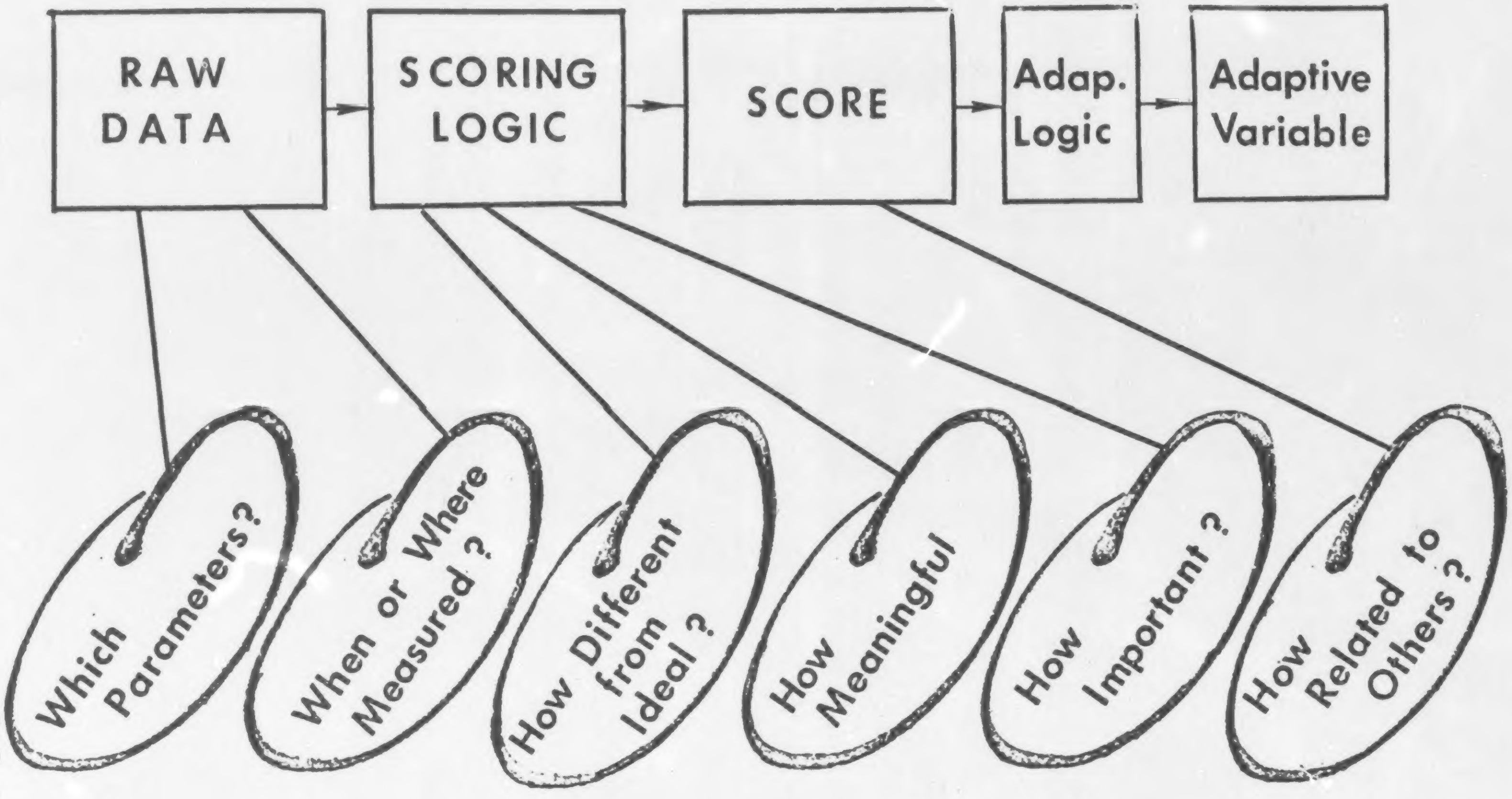

RAW DATA
SCORING LOGIC
SCORE
Adap. Logic
Adaptive Variable
Which Parameters?
When or Where Measured ?
How Different from Ideal ?
How Meaningful
How Important ?
How Related to Others ?

FIG 13

SECTION IV

TEST DEVELOPMENT

USE OF OCCUPATIONAL SURVEY DATA IN DEVELOPING OUTLINES FOR SPECIALTY KNOWLEDGE TESTS

By

Nancy A. Lewis
Senior Review Psychologist

USAF Occupational Measurement Center

I work for the United States Air Force Occupational Measurement Center. The Center is tasked with a dual mission. Two branches develop all tests in support of the Weighted Airman Promotion System, and one branch performs occupational surveys of enlisted career ladders. These two functions, while distinct, are closely related in that both measure jobs of enlisted personnel throughout the Air Force. My presentation concerns the ways in which each function is proving to be a great resource to the other and in particular, the ways in which occupational survey data may be used in developing Specialty Knowledge Test outlines.

An occupational survey for a particular job or Air Force Specialty begins in the Inventory Development Section of the Occupational Survey Branch. An inventory developer coordinates closely with Headquarters USAF, Air Training Command, and the resident technical school for that specialty. He makes TDY trips to visit actual job sites and develops a list of all tasks performed in the specialty. From his research, job inventory booklets are prepared, validated, and administered to individuals who are actually performing jobs within the specialty. These individuals represent every command having personnel assigned to the specialty. Great care is given through statistical sampling monitoring to insure that a truly representative sample is obtained. When the inventory booklets are returned, the data is fed into a UNIVAC dual 1108 computer and is then sent to the Occupational Analysis Section where an analyst assesses commonalities and differences in jobs within the specialty as they are actually being performed throughout the Air Force. His analysis is incorporated into management decisions to determine what changes, if any, should be made regarding career field organization and/or training to be provided for airmen in the specialty.

Our work in the testing function deals primarily with the construction of Specialty Knowledge Tests (SKTs). These tests are designed to test the knowledge necessary to perform the job and are used for promotion to enlisted pay grades E-5, E-6, and E-7 for both active duty and Air National Guard/Reserve personnel. Typically, these SKTs are revised annually. The normal pattern is to perform a Major Revision every 2 years and a Minor Revision or updating of the test on the off year. The revisions are performed by groups of senior NCOs who are assigned throughout the Air Force and who come TDY to the Center expressly for the purpose of providing technical expertise needed for writing appropriate questions.

During a Major Revision, which is normally a 6-week project, these NCOs or Subject-Matter Specialists (SMSs) as we call them, work under the close supervision of a test psychologist trained in the principles of psychometrics and experienced in Specialty Knowledge Test construction. The SMSs are experts currently working in the specialty for which tests are being revised, and they represent the major using commands, i.e., those commands having the greatest number of personnel working in the specialty.

Ever since the testing and survey functions became part of the same organization, these Subject-Matter Specialists have assisted occupational analysts who come to talk to them during the test revision project. The SMSs can often provide insight to the analyst who is looking at the voluminous data which the surveys provide concerning the specialty which he is surveying. The team is at his disposal to exchange ideas, discuss trends, and provide additional "inside" information in great detail. And they have often done so. A Major Revision team provides simultaneously a wide spectrum of knowledge and opinions that cannot easily be sampled by TDY trips or be inferred from the data.

Conversely, since the two functions have merged, a few stalwart test psychologists have seen the merits of using the occupational data as input to the testing program and have gone to great lengths to do so with a large degree of success. But these people were in the minority because of the difficulty of the undertaking. The data provided by a survey is extensive and was not in a readily usable format for test construction purposes. In spite of numerous attempts to train our test psychologists in the applications of survey data, use of the data did not become widespread.

Both the testing people and the survey people were aware that a new format for survey data was needed. A committee was formed to work out a solution. At last, the surveys are now being regularly put to work in the testing area. This is how it is being done:

During a Major Revision, each test for a particular Air Force Specialty is almost completely revised. It is certainly completely restructured. We begin construction for each test with a plan or test outline which the test team constructs. It consists of groupings of those tasks which must be tested to represent a broad, unbiased sample of the examinee's knowledge of the total job he must perform. As a basis for the outline, we use the Air Force Specialty Description in Air Force Manual 39-1, which is a clear-language description of the duties performed in that specialty at each skill level (3-level or apprentice level, 5-level or specialist level, and 7-skill level or technician/supervisor level). We also use the Specialty Training Standard, which is used in the Air Force primarily for on-the-job training. This document is a detailed breakdown of the job and lists in lengthy outline form the tasks in which an apprentice, specialist, or technician/supervisor must attain proficiency. This training standard or STS also indicates the desired level of knowledge/proficiency to be attained at each of these skill levels.

All tasks/duties covered in these two documents must be covered by the test outline. The test team must decide upon logical groupings of these tasks, list them in outline form, and then determine how many questions of the 100 questions on an SKT should be assigned to each outline area breakdown. Outline development is not an easy task, and it has a tremendous effect on the quality of the test which is produced. It requires diligent team effort to arrive at a sound consensus based on individual experience and rationale. We feel that herein lies one of the primary advantages of the Air Force's method of test construction. Decisions are not left to one individual but are arrived at conjointly by individuals who represent long years of experience in various commands and in assignments which call for varied applications of the many-faceted knowledge inherent to a specialty.

Arriving at a consensus can be an exciting, challenging, and enlightening process. Most times, agreements are reached quite easily. But sometimes we discover such things as the fact that one command rarely performs a certain task that the others perform often or the reverse situation, in which one command performs a task that the others do not. Such findings must be taken into consideration in outline development.

Generally, this group approach works very well. However, personalities and command interests understandably enter in. There may be an individual on the team who may try to sway the other members to do things his way. Or there may be several individuals who become argumentative and have difficulty reaching an agreement. The test psychologist has been trained to guard against such occurrences, but it is a difficult matter unless he has not only tact but facts to back him up. Also, there are instances when the team members sincerely wnat to make proper judgments but feel unsure of the proper course of action. In such cases as these, occupational survey data can prove to be invaluable, for surveys provide a more scientific method of making these judgments.

As I mentioned earlier, the committee which consisted of both test personnel and survey personnel has worked together to find a way in which survey data can be readily used for test purposes. We have developed an initial occupational survey data product which will for the interim best serve the requirements of outline development. A special, separately bound computer printout is now being routinely provided for the use of test construction teams. Factors in determining this set of printouts included the size of the package, relative simplicity of use, completeness of data, and analyst time required for obtaining and binding the data. It has been operationally tested by test development personnel and found to be extremely useful.

The first step in outline development is to decide upon logical groupings of the tasks involved and to list them in outline form. Since the occupational survey provides a list of all tasks performed by personnel in the specialty, it can serve as a guide for the team to determine what topics should be covered by the outline and to insure that none is overlooked. The survey itself groups the tasks into several duties. Sometimes these duties can be used to suggest possible outline groupings; however, they should only serve as suggestions in that they may not be the best possible way of grouping the tasks for test purposes. The more complete task list should be the major consideration.

The new printout separates the data into three groups: apprentice level, specialist level, and technician or first-line supervisor level (E-6 and E-7 pay grades combined). This meets our normal requirements because we usually develop an active duty test for each of these levels.

Normally, the technician/supervisor-level outline is developed first. Since the occupational survey provides data by skill level, it can be helpful in determining what topics are appropriate to this level and to distinguish differences between tasks performed at this level and those performed at a lower level. In other words, it can be used to establish boundaries regarding the knowledge to be tested at each level.

For maintenance-type specialties, the printout offers an added feature. It lists all equipment operated/maintained by personnel in the specialty at each level. This summary is quite helpful in determining outline topics. We would certainly not want to ask questions on equipment that few members within the specialty work with; nor would we want to ask questions at the specialist or apprentice level when only technician-level personnel work with the equipment.

After outline groupings have been chosen and listed, the next step is to determine how many questions should be assigned to each outline breakdown. It is in apportioning percentage weights to the outline that our computer printout becomes most useful. Several variables must be taken into account. Our printout provides data regarding these variables. For each of the three levels, the printout offers the following categories of information relative to each task performed by personnel in the specialty:

(1) Percent time spent - This is the relative amount of time the average member working in the specialty devotes to a particular task. It is a good indicator of how often he will be called upon to use the knowledge required for one task as opposed to another.

(2) Percent members performing - This is the percent of those inventoried who are actually called upon to perform a particular task. Logically, more questions should be asked concerning tasks which most

members in the specialty Air-Force-wide perform than those which few members are called on to perform.

(3) Task difficulty - This variable involves the nature of the task itself. Each task is scale-rated by personnel experienced in the specialty as to the amount of time required to learn the task. This variable is a must for consideration, for although a task may be performed by all members and takes a great deal of their time, it may be a fairly routine one such as filing or pushing plainly marked buttons. Questions should be written with regard to the relative difficulty of the task knowledge required.

These three variables have always been taken into consideration when apportioning outline weights, but this printout enables the process to become a more scientific and simple operation.

Although it is only an initial step, we are quite proud of our present printout; for test development purposes, this product is certainly an improvement over the printouts we formerly received. But it is only the beginning. We are still researching ways to improve on this product.

We are considering the feasibility of providing data on other variables. One such variable concerns the consequences of inadequate performance. For example, a task may be easy to learn and rarely performed but may have disastrous results if not performed correctly. This would surely have a bearing on its importance testwise.

One research team is studying mathematically the value of each of these variables as predictors of the way in which a test team would develop an outline. Their long-range goal is to arrive at a single index which would incorporate the variables essential to proper outline development.

We are also weighing the relative merits of alternative formats. One analyst is attempting to devise a system whereby the data contained in the short-version printout could be presented in such a manner that each page relates to a particular Specialty Training Standard paragraph. This would make the printout an extremely simple tool for even the newest test psychologist to use.

I should also mention that in addition to the printout, the extensive backup data which the surveys still provide may also be used to great advantage. For example, when significant command differences are uncovered by the analyst, information is also listed by commands. One can determine exactly which commands perform what tasks and which commands operate and maintain what equipment. By consulting this data, a team can avoid testing on tasks or equipment which are not common to all commands within a specialty and thus perhaps avoid biasing a test for or against a particular group.

The possibilities for using occupational survey data in test outline development are perhaps not infinite, but they are certainly numerous. Some of us even envision the possibility of a computer-developed outline. In these days of trying to economize, a week or so of SMS TDY savings would be welcome!

THE EFFECT OF FORMAT CHANGE ON TEST ITEM DIFFICULTY

Joseph L. Boyd, Jr.
Educational Testing Service

There has been relatively little investigation of the effect that changing the format of multiple choice test items has on item difficulty. Such studies that have been done have dealt primarily with the effects of violating some of the ancient principles of item construction, or on the effect of using three, four, or five answer choices. Dunn, Goldstein, and Berkhouse (1956) studied the effects on difficulty, validity, and reliability of tests by violation of certain item writing principles: use of incomplete statement stems, avoidance of specific determiners, responses of equal length, and consistency of grammar between stem and responses. They found that items with specific determiners, with longer correct choices, and inconsistent grammar clues were easier, but no less reliable or valid, than items "correctly" written.

McMorris, Brown, Snyder, and Pruzek (1972) in a similar study, found that faults made the items easier, but that validity and reliability coefficients were virtually unchanged.

Board and Whitney (1972) found that violating good item writing practices generally made tests easier, as well as less reliable and valid. These, and similar studies, conform to the standard pattern of experimental design.

The study reported here evolved from a practical field situation that dictated an item format change. The item type in question is a variety of the multiple true-false item, widely used in state and municipal civil service examinations. Items are produced, not by combining pairs of independent true-false items, but as a means of salvaging not quite adequate four-choice multiple choice items. The original items may have no right answer; two, three, or four right answers; or one or two ambiguous or non-plausible responses. Conventionally, the stem and two viable responses are put into a format similar to the following:

Example 1. On a truck with a vacuum power booster, too much pedal pressure is needed to apply the brakes. What is a likely cause?

I. An air leak on the brake pedal side of the power booster cylinder

II. A leak in the diaphragm of the power brake booster

(A) I only (B) II only (C) either I or II (D) neither I nor II

Items in this format were developed for use in a national testing program for automobile mechanics. Raw four-choice items were written by instructors and service engineers who had had little item writing experience. Test specialists revised the items, which were then reviewed by separate panels of journeymen mechanics and service engineers. Many items were determined to be faulty, having no or multiple answers, or other ambiguities. The only feasible procedure was to recast the items into the format shown above.

After the items had been used in operational tests, there were protests from both mechanics and their employers that the items were too formal, too artificial, and too hard to read.

A format variation was developed, which seemed to be more clear cut, informal, and easier to read and understand. Item 2, below, is an example of this format.

Example 2. On a truck with a vacuum power booster, too much pedal pressure is needed to apply the brakes.

> Mechanic A says that an air leak on the brake pedal side of the power booster cylinder could be the cause.
>
> Mechanic B says that a leak in the diaphragm of the power brake booster could be the cause.

Who is right?
(A) A only (B) B only (C) both A and B (D) neither A nor B

The revised items were used in the test battery given in the Spring of 1974.

One hundred and four items, distributed among seven of the eight tests of the battery, were identified as having been used in both formats. The item difficulty levels were reviewed. Fifty-eight revised items were more difficult, 36 were easier, and 10 did not change in difficulty. Expressed in terms of percent correct, the mean changes in difficulty of the modified items were as follows:

Test	1	2	3	4	5	6	7
% Mean Change	-0.4	-0.1	0.0	+4.7	-1.3	-1.5	-0.1

For the 104 items, the mean difficulty level change was less than +.03%.

In practical terms, the overall effect of changing item format was to make the test items easier by an amount that would make mean percent correct scores higher by less than one percent. The real effect of the format change has been to almost completely eliminate the protests about "those damn confusing, unreadable test questions."

REFERENCES

Board, C. and Whitney, D. R. "The Effect of Selected Poor Item-Writing Practices on Test Difficulty, Reliability and Validity." Journal of Educational Measurement, 1972, 9, 225-233

Dunn, T. F. et al. "Effect of Item Construction Principles on Difficulty, Reliability, and Validity." PRB Technical Research Note 64, November 1956, Personnel Research Branch, TAGO, DA.

McMorris, R. F. et al. "Effects of Violating Item Construction Principles." Journal of Educational Measurement, 1972, 9 287-295

A 1" x 3½" KEYHOLE ON THE WORLD OF MOS TESTING

Arthur G. Hermansen and Clay V. Brittain

US Army Enlisted Evaluation Center

The effective management of a large military proficiency testing program requires feedback information from the field. Feedback at the Army Enlisted Evaluation Center includes letters from soldiers, telephone calls from Test Control Officers, and occasional nudges from higher headquarters. Feedback from the field can be negative: "the tests are too long, they are too difficult, there are too many unnecessarily difficult words, the questions do not apply to the soldier's job, study references are not available." The list of complaints is a long one.

The responses may take the form of patient explanation of English usage, a short course in statistics, or apologies and promises to do better. What ever the response, one may find oneself in a quandary: one wants to believe that test preparation has been good; that one's item writers are indeed expert; that one's editors have been diligent and professional; that one's tests are readable; that they are really relevant to the soldier's MOS. But in wanting to believe these things, one wonders if they are true. More than likely, one would be hard pressed for empirical evidence that they are.

A useful procedure in such situations is to obtain systematic feedback from the people who take the tests. This paper describes a simple means used by the US Army Enlisted Evaluation Center to get feedback information from soldiers who take our tests. The aim has been to obtain information useful in the quality control of MOS tests and in certain aspects of the testing program.

Beginning with the tests developed for the May 1973 test period, we have included 8 to 10 survey type questions in the test booklet. These follow the test questions and are in a multiple choice format. Earlier, we had revised our standard test answer sheet to include a special management information section. This provides space for mark sensed responses to 10 multiple choice questions. The space measures 1" by 3½"; hence, the whimsical title of this paper.

We, at the Enlisted Evaluation Center, had a number of questions about our tests which we wanted to put to soldiers who take these tests: How does the soldier see the test in terms of difficulty, relevance, readability, length, appearance, etc? We asked such questions, but we also found that managers at the proponent agencies which have responsibility for developing the MOS tests had questions of their own which they wanted to include in the survey section of the test booklet. Examples of uses of the survey data by Item Writing Agencies follow:

The US Army Military Police School issued consolidated study references for selected MOS tests. The school submitted survey questions to be used in determining whether soldiers were able to obtain the study references and how useful they were found to be. These questions appeared in MOS 95B test booklets for the August 1974 test period and are scheduled to appear in the MOS 95C test booklets for the February 1975 test period.

The US Army Infantry School has included in certain tests an experimental section called "Operations in a Combat Environment." The experimental section represents an attempt to phrase test questions in terms of realistic situations. As a means of evaluating the success of this effort, the Infantry School submitted five survey questions to be included in the test booklets. These questions were designed to determine whether soldiers actually found the experimental section of the test more realistic and more effective as a means of testing MOS proficiency.

Although the added time requirement is slight, the survey questions do add to the testing burden imposed upon the soldier. The soldier who has spent two, three, or even four hours taking an MOS test may not be happy when required to answer an additional 10 questions which are not really part of the test. Out of deference to this concern, soldiers' responses to the survey questions have not been made mandatory. That is, Test Control Officers have been instructed to inform soldiers that the questions are to be answered on a voluntary basis. The overall response rate has been about 50 percent. This, of course, introduces the problem of selectivity. What type of soldiers respond to the survey questions and how do they differ from non-respondents? We do not know the answer to this question but we do know that, in terms of test score, respondents do not differ consistently from non-respondents. We are not getting responses simply from those who do well on the tests.

Following are examples of questions which have been included in recent surveys and the responses to them:

SLIDE 1
SLIDE 2
SLIDE 3

In addition to reporting aggregate data, we have the capability of breaking down and reporting responses by specific MOSC or MOS; hence, we can provide component agencies with information which is directly relevant to their tests.

I might note in closing that data from the survey questions have served at times to boost our morale. In one survey, we found that the soldier who complains about undue length of the test is in a minority. Seventeen percent of the soldiers said the tests were too long, but 49 percent thought that they were about right...and 5 percent thought they were too short. The others had no opinion in this regard. In another case, we found that only 10 percent of the soldiers felt that the tests were unnecessarily difficult to read.

But such responses are not the major aim of the survey questions. The purpose is to obtain information which is relevant to the management and improvement of the MOS testing program. The questions are a convenient and inexpensive means of getting such information. Responding to the questions probably takes less than 10 minutes of the soldier's time, and the responses can be processed at the same time that the tests are machine scored. We believe this ONE INCH BY THREE-AND-A-INCH KEYHOLE ON THE WORLD OF MOS TESTING is useful to us and to the agencies with which we work.

SLIDE 1

4. Question: Which statement describes BEST your feelings toward the test? (A-7)

ALTERNATIVE	MAY & AUG 73 RESPONSE	MAY & AUG 73 %	NOV 73 RESPONSE	NOV 73 %
A. The questions concern things I need to know to be fully qualified in my PMOS	8,989	56	9,788	42
B. Some of the questions are not related to my PMOS	4,616	29	6,566	28
C. Most of the questions are not related to my PMOS	1,383	8	2,738	11
D. Very few of the questions are related to my PMOS	1,116	7	4,367	19
TOTAL	16,104	100	23,459	100

COMMENT: It appears as if a problem area is beginning to appear in the area of developing MOS-related questions, as reflected in the increase of C and D responses. Since alternative D reflects the strongest negative reaction, the scores of those soldiers selecting D were analyzed - and it's true. Of the 4,367 soldiers who selected D, 3,716 or 85% received lower than mean scores for their MOS.

SLIDE 2

17. Question: Compared with the last test I took in this MOS, this test is (C-3)

ALTERNATIVE	MAY & AUG 73 RESPONSE	MAY & AUG 73 %	NOV 73 RESPONSE	NOV 73 %
A. a lot better	1,760	14	2,918	14
B. somewhat better	2,763	22	4,366	21
C. about the same	6,109	49	9,726	48
D. not as good	1,851	15	3,435	17
TOTAL	12,483	100	20,445	100

SLIDE 3

20. Question: Which statement describes BEST your experience in getting the references listed in the study guide? (C-6)

	ALTERNATIVE	MAY & AUG 73 RESPONSE	%	NOV 73 RESPONSE	%
A.	I got them all	2,728	21	3,547	16
B.	I got them all but I could not keep them long enough	3,341	25	4,998	23
C.	I could not find them all and nobody would help me get them	3,416	26	5,520	25
D.	I did not really try to locate them all	3,685	28	7,839	36
	TOTAL	13,170	100	21,904	100

COMMENT: Response to alternative A has acquired direction over the past three survey periods. In May 1973, 22%; in August 1973, 20%; and in November 1973, 16% of the soldiers were able to find all of the study materials. Do our MOS librariers have sufficient numbers of the required publications on hand? In general, those soldiers who found all of the study materials fared well. Of those who selected the A response, 3,218 or 9.07% achieved higher than mean scores on their MOS test.

SECTION V

NEW TEST TYPES AND TEST VALIDATION

A NEW SEQUENTIAL MULTIPLE-CHOICE SKILL-RELATED TESTING DEVICE

1 October 1974

F. B. Braun
J. E. Tindall, Jr.

Data-Design Laboratories
Norfolk, Virginia 23502

A NEW SEQUENTIAL MULTIPLE-CHOICE SKILL-RELATED TESTING DEVICE

INTRODUCTION

This paper discusses the development of a new testing device designed to meet specific testing requirements for the Fleet Ballistic Missile Weapons System (FBMWS) Personnel and Training Evaluation Program (PTEP). Project development was fostered by Data-Design Laboratories under contract with the Navy Strategic Systems Project Office, and was coordinated through the Central Test Site for PTEP at the Naval Guided Missiles School, Dam Neck, Virginia Beach, Virginia.

BACKGROUND

PTEP is tasked with performing the qualitative assessment of the knowledge and skill levels of FBMWS personnel while undergoing formal training and while assigned to FBM submarines, tenders, and training facilities. PTEP is also assigned the responsibility of conducting evaluations of training facilities, hardware, documentation, and instructional materials as a basis for implementing improvements in the training system. The largest input of evaluation data is collected through the periodic administration of achievement tests to FBMWS personnel. The testing population is approximately 4,000 technicians, the majority of whom are tested twice a year.

The PTEP achievement tests employ two types of test instruments. Conventional, four-alternative multiple-choice test items are used to measure skill-supporting theory; and unique pencil-and-paper exercises are used to measure skill performance in terms of skill-related knowledge application. These exercises (referred to as "skill exercises") employ realistic simulations of equipment and the use of authorized procedures for equipment operation and maintenance. It is the development of these skill exercises that is the subject of this discussion.

DEVELOPMENT OF THE NEW SKILL EXERCISE

PTEP has had considerable experience with different types of skill-related testing devices. The pencil-and-paper means of skill-related testing has been determined to be particularly well suited for FBMWS PTEP applications for two obvious reasons. First, the expense of FBMWS training hardware prohibits its dedicated use in PTEP to the extent necessary for adequate and effective personnel achievement testing; and second, the FBMWS equipment is largely electronic and therefore requires few psychomotor skills.

Less obvious reasons for selection of the skill exercise over "hands-on" performance testing in PTEP are the higher degree of specificity with which the skill exercise can identify skill deficiencies, and the scoring objectivity through which those deficiencies can be consistently identified. Although some performance tests may provide for this objectivity and specificity, most do not.

DESIGN OBJECTIVES. It is based on this PTEP experience that the effort was undertaken to establish the design objectives for an improved testing vehicle tailored to the needs of PTEP achievement testing. Certainly, those objectives are not unlike those of many instructional measurement systems. The following summarizes the original design objectives.

1. Validity. The test results should correlate well with actual on-the-equipment performance.
    - a. Realism. During the test procedure, the examinee should have ready reference to equipment front panel indications, waveforms, voltage measurements, machine configurations, etc.
    - b. Objectivity. Test results should be independent of grader subjectivity and error.
    - c. Reliability. Test results should be consistent for multi-measurements.
    - d. Relevancy. Test instrument design should be keyed to task analysis.
2. Efficiency. The entire personnel testing process should be critically conservative of examinee and test administrator requirements.
    - a. Output. The test instrument should provide greatest number of response data per unit of examinee time to enhance evaluative data base.

b. Administration. Administration of the test instrument should require minimum examinee time and test administrator participation.

3. Economy. The test instrument should be constructed in the most cost-effective manner.
   a. Production. The physical characteristics of the test instrument should minimize printing and binding costs.
   b. Handling. The amount of support information integrated into the test instrument should produce minimum test material bulk to reduce storage and shipping costs.
   c. Preparation. Data collection and test construction techniques should be standardized to minimize test instrument design costs.

One of the most important design considerations was to eliminate the scoring problems encountered with previous skill exercises. These are problems which stemmed from attempting to standardize and automate scoring of exercises covering diverse topics, each of which permitted several possible intra-exercise solution paths from which the examinee could select.

After consideration of all desired design objectives, the following decisions were made, upon which preparation of the prototype exercise was based.

1. Use multiple-choice items for the exercise building blocks, thereby facilitating objective scoring and results specificity.
2. Sequence the items to develop a single-path exercise solution, thereby facilitating a uniform scoring methodology.
3. Employ latent-image techniques for solution masking and response feedback, thereby reducing production costs with respect to erasable-ink overlay processes.
4. Integrate latent-image technique with mark-sense format to enable machine processing of raw test data, thereby minimizing test administrator involvement and scoring errors and, as well, more fully utilizing available mark-sense hardware currently used for knowledge test item processing.

EXERCISE FORMAT. The skill exercise is formatted to present the examinee with a realistic situation of equipment operation or maintenance activity, and then requires him to solve a procedural problem involved with that activity. The examinee solves the exercise by identifying the correct procedural steps he should normally take, considering the alternatives presented, to perform the particular equipment operation or to remedy the apparent equipment malfunction. Although each exercise may not necessarily reflect the exact procedure which the examinee would prefer to use, it is expected that a trained technician would recognize the exercise solution path as being valid and the most appropriate one from among those choices made available to him. His ability to recognize and logically work through the exercise solution path is considered to be directly indicative of his ability to apply the various diagnostic skills required of him when encountering the several different possible situations of equipment operation and maintenance. However, specific correlation cannot be determined until external validation studies are completed.

The particular procedural steps which make-up the exercise solution are selected as necessary to fulfill the defined testing objectives of the exercise with respect to testing specific skills normally required of the examinee.

The exercise solution may involve individual steps such as the following.

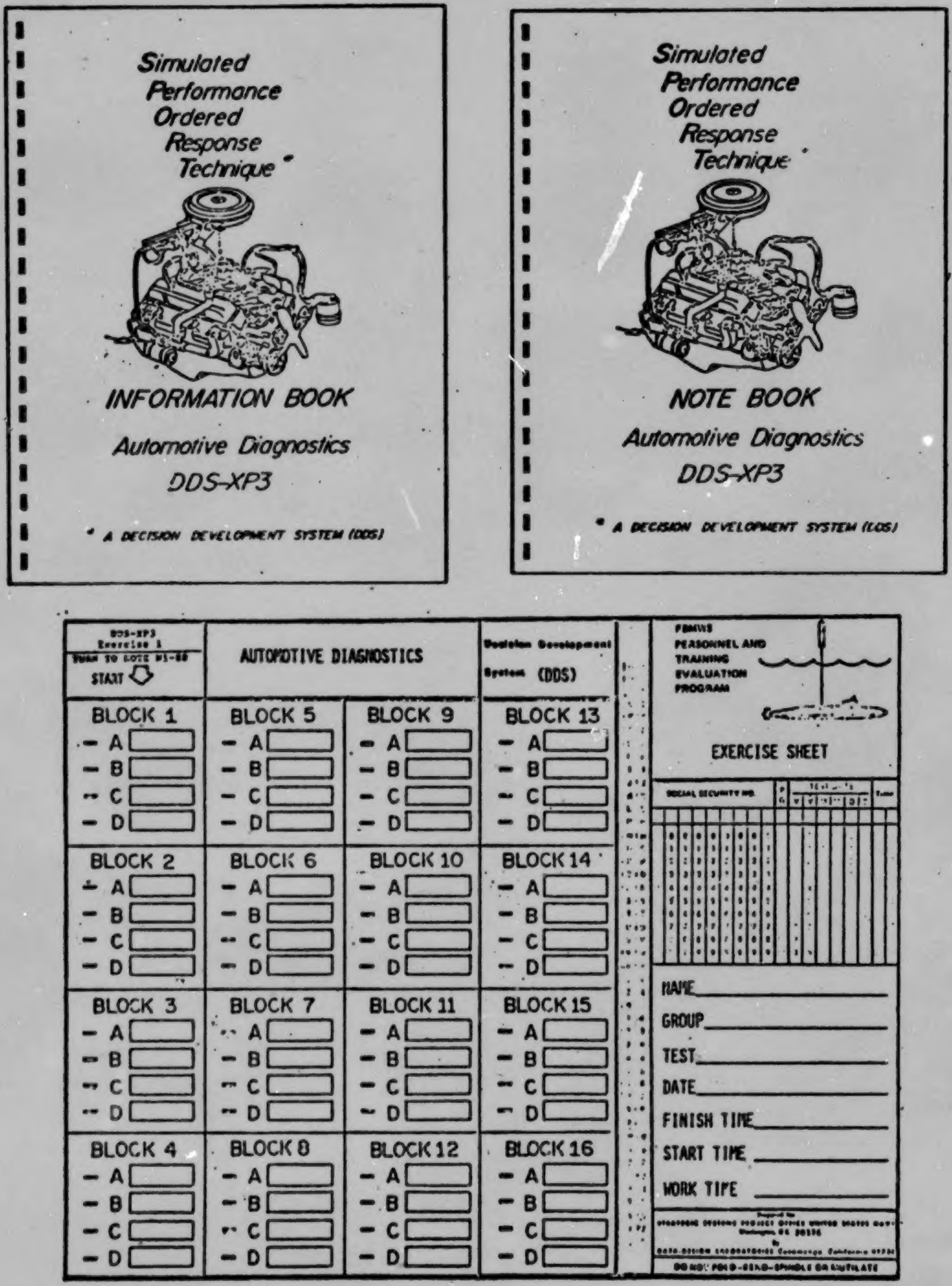
1. Positioning controls in appropriate sequence.
2. Recognition of malfunction indications.
3. Interpretation of equipment printouts and signal waveforms.
4. Isolation and location of equipment faults.
5. Selection and identification of test points within the equipment to be monitored.
6. Location and identification of appropriate documented procedures to be used.
7. Selection of logical improvised procedures as authorized.
8. Observation of safety precautions.

The exercise solution is supported by realistic pictorials of control panel indications, duplications of actual equipment printouts, photographs of waveforms as observed with an oscilloscope, and the same documented procedures, technical data, and equipment maintenance diagrams as would normally be available in the real environment of equipment operation and maintenance.

Exercise Components. For the most cost-effective design, preparation, and production, each skill exercise is packaged within a set of other subject-related exercises. A skill exercise set consists of the following components. (Figure 1 shows the components of the prototype skill exercise set.)

1. Information Book. This booklet contains general directions to the examinee and the pictorials, printouts, waveforms, etc. which are necessary to the solution of the exercises. These are the indications which normally would be used to interpret equipment conditions during the performance of the procedures portrayed in the exercise solutions.
2. Note Book. This booklet contains the individual multiple-choice test items (referred to as "notes") which collectively form the procedural test for the exercises. Each note represents a separate exercise building block (or step) in the exercise solution path.
3. Exercise Sheets. Each exercise has a separate sheet on which the examinee marks his decision for each step of the exercise solution (i.e., alternative selection for each note), and from which the examinee receives an immediate response to each of his selections indicating correct or incorrect. If a selection is incorrect, the examinee is directed to review the same note and select another alternative. If a selection is correct, the Exercise Sheet provides the examinee with a reference to the next note in the exercise solution sequence; and he then evaluates that note in the Note Book and makes his next selection on the Exercise Sheet from among the new set of four alternatives provided.

   The selection responses (incorrect-choice indicators and correct-choice note references) are printed on the Exercise



Simulated Performance Ordered Response Technique*

INFORMATION BOOK

Automotive Diagnostics

DDS-XP3

* A DECISION DEVELOPMENT SYSTEM (DDS)

Simulated Performance Ordered Response Technique*

NOTE BOOK

Automotive Diagnostics

DDS-XP3

* A DECISION DEVELOPMENT SYSTEM (DDS)

DDS-XP3
Exercise 1

START

AUTOMOTIVE DIAGNOSTICS

Decision Development System (DDS)

BLOCK 1	BLOCK 5	BLOCK 9	BLOCK 13
– A	– A	– A	– A
– B	– B	– B	– B
– C	– C	– C	– C
– D	– D	– D	– D
BLOCK 2	BLOCK 6	BLOCK 10	BLOCK 14
– A	– A	– A	– A
– B	– B	– B	– B
– C	– C	– C	– C
– D	– D	– D	– D
BLOCK 3	BLOCK 7	BLOCK 11	BLOCK 15
– A	– A	– A	– A
– B	– B	– B	– B
– C	– C	– C	– C
– D	– D	– D	– D
BLOCK 4	BLOCK 8	BLOCK 12	BLOCK 16
– A	– A	– A	– A
– B	– B	– B	– B
– C	– C	– C	– C
– D	– D	– D	– D

PERSONNEL AND TRAINING EVALUATION PROGRAM

EXERCISE SHEET

SOCIAL SECURITY NO.

NAME__________

GROUP__________

TEST__________

DATE__________

FINISH TIME__________

START TIME__________

WORK TIME__________

Figure 1. Prototype Skill Exercise Components

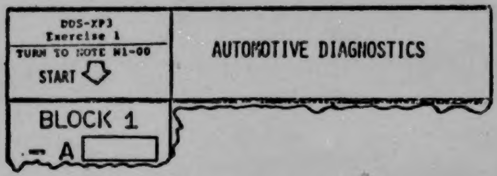
Sheet in latent-image ink, and are revealed only when the examinee swipes the appropriate areas with the special marking pen provided. Thus, a permanent, visible record of all selections made is created as the examinee proceeds through the exercise solution.

Mark-sense areas are also defined on the Exercise Sheet for the examinee to mark, corresponding to each selection made. This provides for automatic machine-reading of the specific solution path developed by the examinee in working the exercise, and serves as the raw test data input to the automatic scoring process.

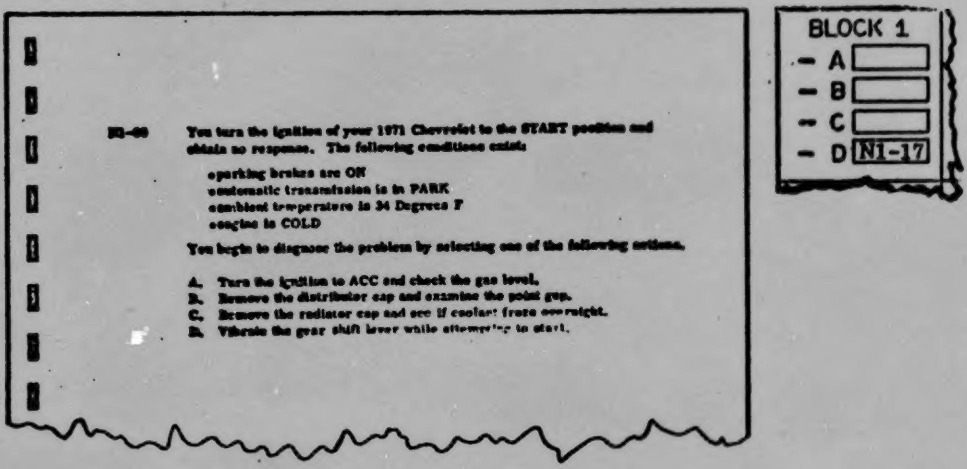
Exercise Example. To demonstrate the new approach, the prototype skill exercise was prepared. This exercise depicts the device methodology, rather than to provide an actual exercise with which personnel skills can be accurately measured.

The prototype skill exercise was built on a familiar piece of equipment: the automobile. This exercise has been submitted to several FBMWS training system participants and other training specialists to solicit constructive criticism of the new approach. A brief description of the prototype exercise is provided in the following paragraphs.

1. The exercise begins with a note reference on the Exercise Sheet.



2. Turning to Note N1-00, the examinee is informed of the initial conditions, and then proceeds to diagnose the problem.



N1-00 You turn the ignition of your 1971 Chevrolet to the START position and obtain no response. The following conditions exist:

- parking brakes are ON
- automatic transmission is in PARK
- ambient temperature is 34 Degrees F
- engine is COLD

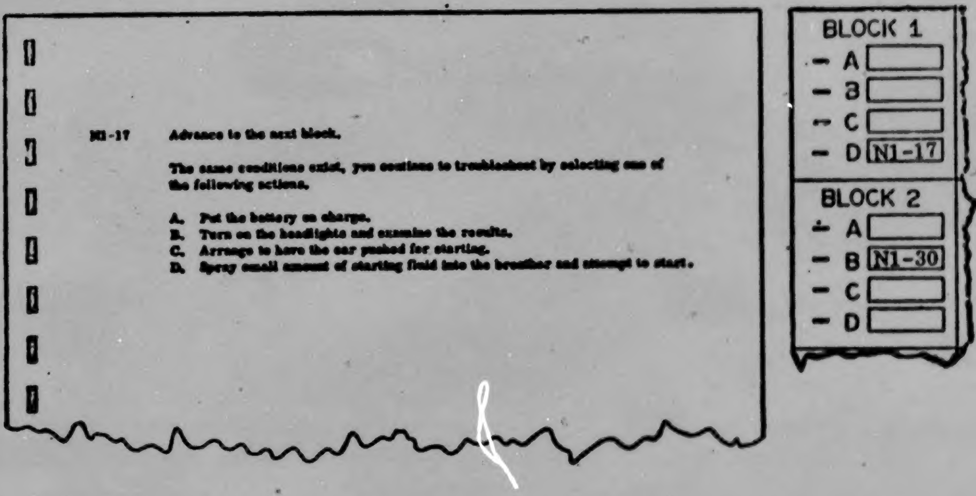
You begin to diagnose the problem by selecting one of the following actions.

A. Turn the ignition to ACC and check the gas level.
B. Remove the distributor cap and examine the point gap.
C. Remove the radiator cap and see if coolant froze overnight.
D. Vibrate the gear shift lever while attempting to start.

BLOCK 1
- A
- B
- C
- D N1-17

Of the alternatives provided, "D" is the most logical choice, because it is apparent to the technician that the problem is more than likely within the electrical starting system, and not in the fuel, cooling, or firing systems. Therefore, the examinee swipes the selection area adjacent to "D" in Block 1 for his response. In a few seconds, another note reference appears: "N1-17". He then proceeds to Note N1-17.

3. This was the correct selection, because the examinee is instructed to proceed to the next block (Block 2) for his next response selection. (Otherwise, he would have been told to make another selection from the same block.)



N1-17 Advance to the next block.

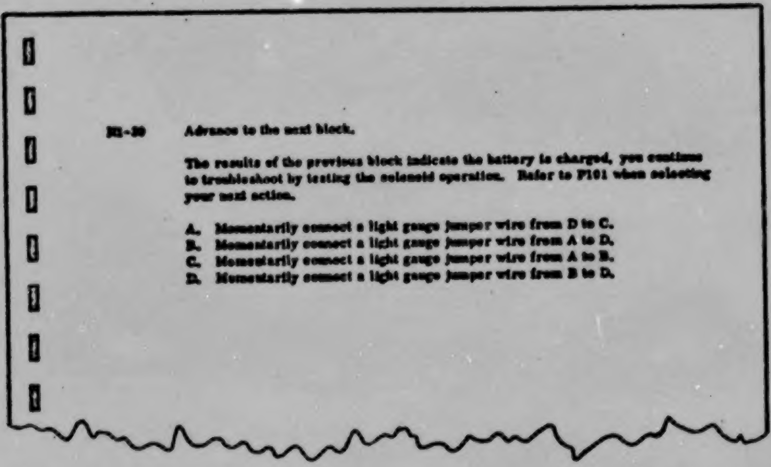
The same conditions exist, you continue to troubleshoot by selecting one of the following actions.

A. Put the battery on charge.
B. Turn on the headlights and examine the results.
C. Arrange to have the car pushed for starting.
D. Spray small amount of starting fluid into the breather and attempt to start.

BLOCK 1
- A
- B
- C
- D N1-17

BLOCK 2
- A
- B N1-30
- C
- D

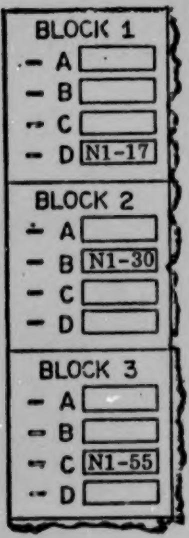
Here, he determines that alternative "B" is correct. (One would not charge the battery until determining whether the battery is low, such as by checking headlight operation.) Similarly as before, he swipes selection area "B" in Block 2 for further development.

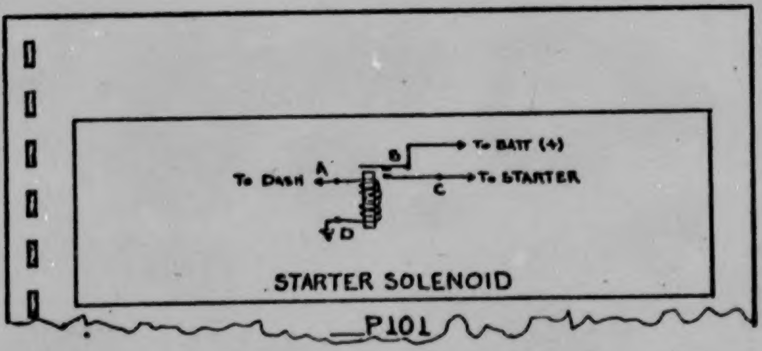


N1-30 Advance to the next block.

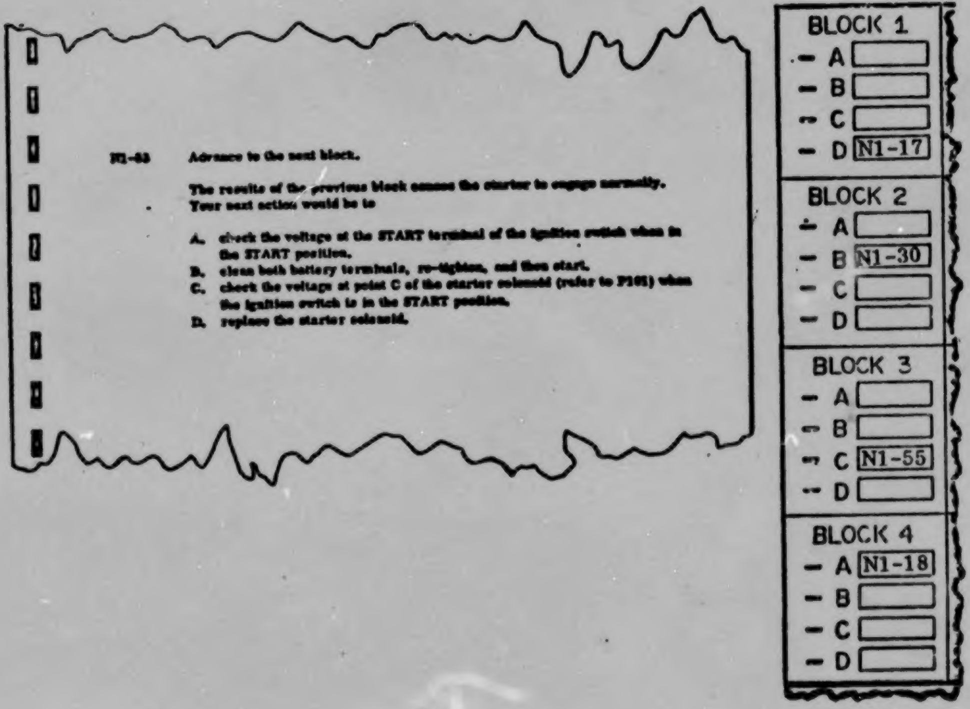
The results of the previous block indicate the battery is charged, you continue to troubleshoot by testing the solenoid operation. Refer to P101 when selecting your next action.

A. Momentarily connect a light gauge jumper wire from D to C.
B. Momentarily connect a light gauge jumper wire from A to D.
C. Momentarily connect a light gauge jumper wire from A to B.
D. Momentarily connect a light gauge jumper wire from B to D.


BLOCK 1
A
B
C
D N1-17
BLOCK 2
A
B N1-30
C
D
BLOCK 3
A
B
C N1-55
D


To BATT (+)
B
A
To DASH
To STARTER
C
D
STARTER SOLENOID
P101

4. This block identifies that the battery is good and the examinee continues to diagnose the electrical starting system by checking solenoid operation. He is referenced to the Information Book for a solenoid schematic (P101).

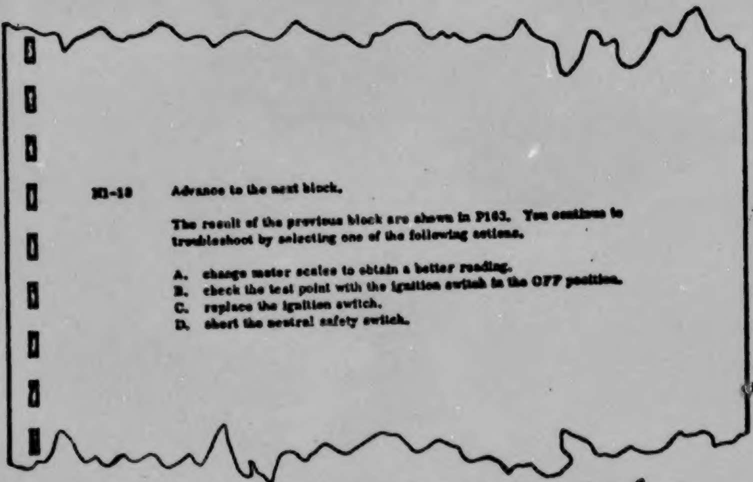


N1-55 Advance to the next block.

The results of the previous block assume the starter to engage normally. Your next action would be to

A. check the voltage at the START terminal of the ignition switch when in the START position.
B. clean both battery terminals, re-tighten, and then start.
C. check the voltage at point C of the starter solenoid (refer to P101) when the ignition switch is in the START position.
D. replace the starter solenoid.

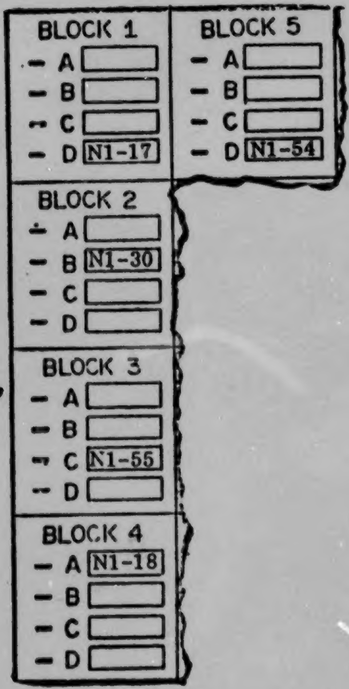
5. The examinee now knows that the battery, starter, and starter solenoid are performing satisfactorily. Therefore, the problem must be in the control circuitry; and he continues by checking ignition switch operation.



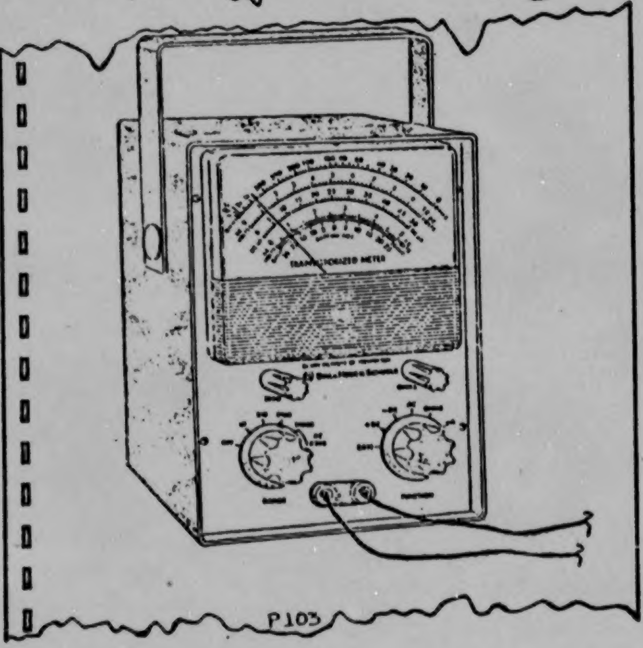
N1-18 **Advance to the next block.**

The result of the previous block are shown in P103. You continue to troubleshoot by selecting one of the following actions.

A. change meter scales to obtain a better reading.
B. check the test point with the ignition switch in the OFF position.
C. replace the ignition switch.
D. short the neutral safety switch.

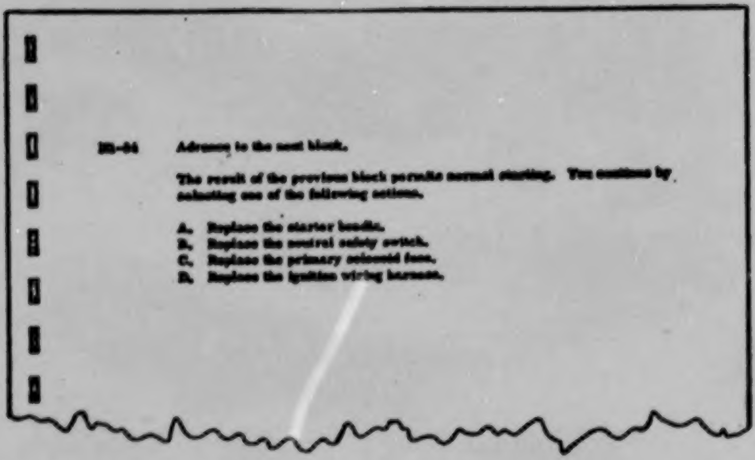


	A	B	C	D
BLOCK 1				N1-17
BLOCK 2		N1-30		
BLOCK 3			N1-55	
BLOCK 4	N1-18			
BLOCK 5				N1-54



P103

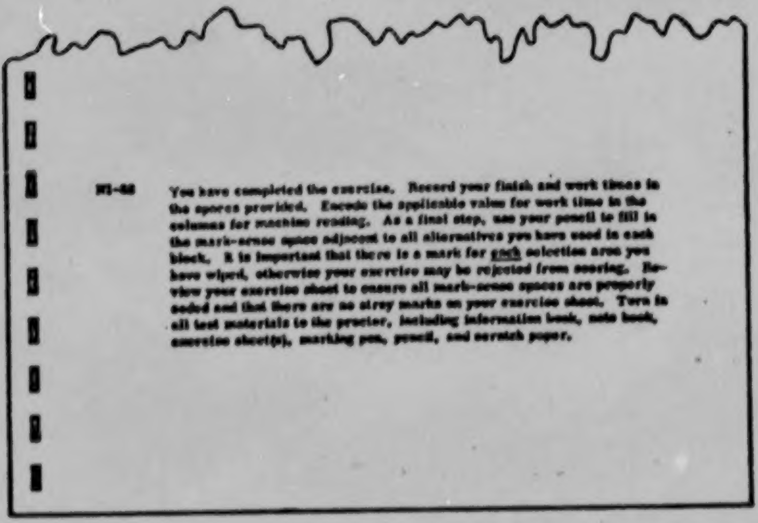
6. This block tests the examinee's understanding of volt-ohm meter test equipment and starter control circuitry. For example, selection of alternative "A" would (in a 'hands-on" situation) "peg" the meter.



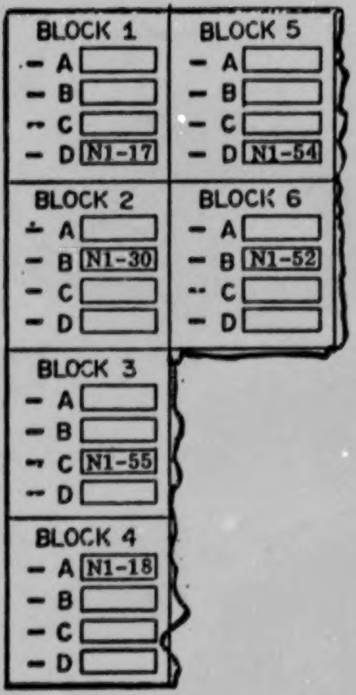
N1-54 Advance to the next block.

The result of the previous block permits normal starting. You continue by selecting one of the following actions.

A. Replace the starter bendix.
B. Replace the neutral safety switch.
C. Replace the primary solenoid fuse.
D. Replace the ignition wiring harness.



N1-52 You have completed the exercise. Record your finish and work times in the spaces provided. Encode the applicable value for work time in the columns for machine reading. As a final step, use your pencil to fill in the mark-sense space adjacent to all alternatives you have used in each block. It is important that there is a mark for each selection area you have wiped, otherwise your exercise may be rejected from scoring. Review your exercise sheet to ensure all mark-sense spaces are properly coded and that there are no stray marks on your exercise sheet. Turn in all test materials to the proctor, including information book, note book, exercise sheet(s), marking pen, pencil, and scratch paper.



BLOCK 1	BLOCK 5
— A	— A
— B	— B
— C	— C
— D N1-17	— D N1-54
BLOCK 2	BLOCK 6
— A	— A
— B N1-30	— B N1-52
— C	— C
— D	— D
BLOCK 3	
— A	
— B	
— C N1-55	
— D	
BLOCK 4	
— A N1-18	
— B	
— C	
— D	

7. To the experienced technician, "B" is the only valid choice; the neutral safety switch is apparently defective.

EXERCISE SCORING. Following the prototype stage, skill exercises were prepared on FBMWS equipment and were formally administered in PTEP tests to obtain actual performance data from FBMWS technicians. Those test results were used to assist in the development of the exercise scoring technique. Since similar on-the-equipment performance data were not available, recent PTEP achievement test results were used as the initial validating criteria.

Many scoring schemes were considered for the new device. The majority were designed around various block-level raw score assignments and some treatment of completion time as primary scoring factors.

The scoring system which was adapted yields examinee performance at the "block" level; that is, information is recorded for each discrete logic portion. An examinee is awarded points based on the number of attempts he makes within a block before advancing to the next block. (An examinee cannot advance to the next block until a correct response is given.) The block-level point distribution was developed based on the intra-block guessing probability distribution. A 0- to 12- point distribution was selected, and then corrected for guessing. Zero, minimum points, was established to avoid processing with negative values. Twelve points was selected because it would provide integer values when the point distribution was corrected for guessing. The following table shows block level score assignments.

	Linear Point Distribution	Guessing Probability Distribution	Guessing Corrected Point Distribution
One Attempt	12	.25	9
Two Attempts	8	.33	8
Three Attempts	4	.50	6
Four Attempts	0	1.00	0

These block-level, guessing-corrected point values are summed to generate the overall exercise raw score, as follows:

$$S_R = \sum_{i=1}^{n} b_i$$

Where:

S_R = exercise raw score

n = no. of blocks in the exercise

b_i = block-level score for the i^{th} block

To assist in interpreting the exercise raw score it is necessary to know the maximum score (S_m) and the guessing score (S_g). The maximum raw score for any exercise is simply 9·n. The guessing score may be calculated as follows:

$$S_g = \frac{n}{4} \cdot 9 + \frac{n}{4} \cdot 8 + \frac{n}{4} \cdot 6$$

$$S_g = 5.75n$$

Where:

S_g = guessing score

n = no. of blocks in the exercise

To simplify test results interpretation the raw exercise scores are subjected to a linear transformation. This facilitates inter-exercise score comparisons for exercises of different block lengths. This transformation yields scores from 0 to 100, with a guessing score equivalent to 25 and the maximum score equivalent to 100, regardless of the number of exercise blocks. Negative scores are set to zero. The following formula applies.

$$S_T = \frac{23.0769}{n} \cdot S_R - 107.692$$

Where:

S_T = transformed score

n = no. of blocks in the exercise

S_R = raw exercise score

Note that this transformation provides scores analogous to four-alternative, multiple-choice percent scores.

This scoring system does not include exercise completion time as a scoring parameter, because attempts to do so degraded correlational results. However, time is recorded for each examinee and may be used in subsequent exercise results evaluation. Regardless, each exercise has a time limit which may not be exceeded. When this time limit is reached the test is terminated. Guessing-level points (5.75) will be awarded examinees for each of the remaining incompleted blocks.

EXERCISE PREPARATION. Each skill exercise is prepared in accordance with defined testing objectives which relate to particular skills required of the technician in his normal role of supporting the FBMWS mission. Subsequent verification is performed to ensure that the exercise fulfills those objectives.

Exercise Design and Construction. Design of the skill exercise is accomplished through close coordination with the constructor of the overall test, of which the skill exercise will become a part. The test constructor specifies the criteria to which the exercise is to be prepared. These criteria are expressed in terms of testing objectives, defining specific skill items to identified levels of proficiency. These skill requirements are determined from the job task analysis delineated for the applicable FBMWS technician in NAVORD OD 43180.

The defined skill requirements are collectively compared with the FBMWS operation and maintenance procedures for which the particular technician is assigned responsibility. The one procedure or set of procedures to which most of the defined skill requirements apply is selected as the basis for technical content of the skill exercise.

Design and layout of the exercise solution path is developed to provide testing of each individual skill item in a separate step (block) within that path. This enables determination of examinee performance on each individual skill item through analysis of test results to the exercise block level. Thereby, specific deficiencies can be identified relative to the personnel and training standards defined by the job task analysis delineation.

The general design and layout of the exercise solution path serve to define the requirements for collection of actual FBMWS equipment pictorials, waveform photographs, printout duplications, etc. These are provided in the Information Book for use by the examinee in working the exercise. It is during the collection of these realistic data from the "live" equipment that the details of the solution path are designed and the actual exercise notes are developed. Technical support is provided by personnel of the Central Test Site for PTEP. Their assistance also adds more realism to the exercise solution by introducing procedural techniques actually experienced during the operation and maintenance of the FBMWS equipment.

The stem of each note (test item) is designed to provide a sequential building block of information with respect to the overall procedure reflected by the exercise solution path. The selection alternatives are worded so that only one choice is correct for each note, while the other choices appear plausible with respect to the particular portion of the exercise solution revealed theretofore.

Exercise Technical Verification. Upon completion of the skill exercise manuscript, it is exposed to a thorough verification to check technical accuracy, currency, and relevancy with respect to the actual FBMWS equipment, documentation, and authorized procedures, and to ensure fulfillment of the testing objectives defined for that particular exercise. Verification is conducted by technically cognizant personnel of the Central Test Site for PTEP. The verifiers actually work the exercise in the same manner as would an examinee and with the same support materials as would be available to him.

Comments from this verification process are reviewed by the exercise designer for incorporation into the manuscript. The testing materials then undergo final production and distribution to the Central Test Site for administration in PTEP achievement tests.

STATUS AND FUTURE PLANS

Currently, exercise development has been completed through the prototype and pilot testing phases. Three complete skill exercise sets are now in preparation with fifteen additional exercise sets averaging fifteen exercises each scheduled to be completed this fiscal year. These exercises are applicable to different FBM Weapon System technicians.

Several of the new skill exercises are being used in an independent project being conducted by another training research activity. This project utilizes a combination of skill exercises, equipment simulators, and hands-on performance tests to prescribe and evaluate shipboard training materials.

Additional skill exercise applications pending are the use of the technique in self-study materials and computer based instruction, with emphasis on remedial and mastery testing. In training-oriented applications, the skill exercise can be used to provide immediate training reinforcement when an incorrect alternative is selected.

After full PTEP implementation of the new skill exercises, the knowledge part of the PTEP achievement tests will be evaluated to determine if test size reductions can be made due to the increased specificity of the new skill part. It is anticipated that reductions may be made without sacrificing overall test reliability.

The simplicity and economy of this new testing device merits its consideration for a wide variety of testing/training applications. Although the device is not intended to replace hands-on performance, it is a potentially effective supplement to situations involving relatively minor psychomotor skills and where expensive training hardware coupled with large testing loads exist.

BEHAVIORAL TOXICOLOGY: A NEW FRONTIER FOR BEHAVIORAL TESTING[1]

Thomas C. Tuttle
Westinghouse Behavioral/Safety Center
Columbia, Maryland

Now perhaps more than ever before in our history, American society is being forced into an awareness of its responsibility to the working man and woman. This awareness has, in the views of many, been slow in coming and long overdue. Many separate forces have stimulated this realization including unions, more concerned management, the media, rising insurance and workmen's compensation costs and of course the government. Certainly one of the most important forces which has, in a sense, legislated a conscience into American industry has been the 1970 Occupational Safety and Health Act. Prior to the passage of this legislation, few companies could afford to provide comprehensive health screening programs, since this cost put them at a competitive disadvantage in the marketplace. Now with the legislation, health and safety standards have been set that apply equally to all companies. Competitive advantages now accrue to those companies that are able to meet the standards with minimum costs. Therefore, there is an increased need for economical yet effective screening procedures that are feasible on a large scale basis in order to provide early warning of any health dangers resulting from workplace conditions.

One very important domain covered by the legislation is worker exposure to toxic chemicals. In the civilian sector there are approximately 60 to 80 million workers distributed over 5 million workplaces exposed to between 500,000 and 600,000 chemical substances. Of these substances, over 13,000 are known to be toxic and can cause harm to exposed workers.

What effects do these toxic chemicals have on workers? Consider the following examples. (1) A worker in a textile factory has always been a mild-mannered agreeable individual. However, after a period of exposure to a toxic substance he becomes aggressive, getting into arguments and fights with fellow workers. (2) A second worker under similar conditions develops such severe headaches that he runs headlong into a concrete wall and then stands there, repeatedly beating his head against the wall. (3) Still another worker suffers social embarrassment resulting from the skin absorption of a chemical having an obnoxious odor. After working a week in the factory he goes to church on Sunday and notices that whenever he sits beside someone they move to the other end of the pew, or else get up and leave. Or consider the damage that might be done to his home life by having his wife always complain that the bedsheets smell like rotten eggs. (4) There are cases in which exposure has been linked to impaired worker safety both on and off the job. An example is the case of a worker exposed to a dry cleaning solvent known to impair coordination and create dizziness if exposure levels are high, who leaves the job at lunch time and becomes involved in an automobile accident. (5) Finally, consider the case of an individual who works with a degreasing solvent during the day and then stops for a few beers on the way home. After consuming approximately one quart of beer, his face becomes covered with fiery red blotches, making him appear like a circus clown. Although this may be an effective deterrent to alcohol consumption, it would seem to be a very heavy price to pay just to earn a living.

---

[1] Presented at the Annual Convention of the Military Testing Association, Oklahoma City, Oklahoma, October 22, 1974.

By this time you are probably beginning to ask yourself, what does all this have to do with behavioral testing, particularly behavioral testing in the military? To answer these questions, let me speak for a few moments about the field that has become known as "Behavioral Toxicology." While my remarks derive from ongoing work in the civilian sector, they have relevance for military psychologists and military environments in two areas. The primary relevance at this point is likely to be methodological. In this room are represented some of the most sophisticated developers and users of psychometric techniques in the country. What I am describing represents in my view a new and exciting use of this technology. If a dialogue can be established, psychologists in the civilian sector can profit from your insights and any new developments that result from work in the civilian sector may also benefit those of you in the military environment.

Secondly, where military personnal are exposed to toxic chemicals, this discussion may provide some suggestions for new research possibilities. My guess is that such exposure conditions in the military work environments pose less of a problem than in the civilian sector. However, there are some potential areas that might be explored. Some possible examples are carbon monoxide exposure of military police who man guardhouses at the gates of busy installations; operating room technicians who are exposed to anesthetics during surgery; ordnance specialists who handle or come into contact with chemical munitions; or loaders in tank crews who are confined to an enclosed turret and have to breathe the fumes from expended rounds during firing exercises. Thus the developing field of Behavioral Toxicology would seem to be an area that might interest behavioral scientists in military settings.

Behavioral Toxicology is a term applied to the use of behavioral measures such as objective psychological tests to detect at the earliest possible time, changes that occur in the worker resulting from exposure to toxic chemicals in the workplace. The importance of this area of study stems from the finding that behavioral changes which can be measured by standard psychometric techniques have been shown in some cases to predate observable clinical signs. Thus by using a properly validated battery of psychological tests, it may be possible to identify workers who are "poisoned" at an earlier stage than they could be diagnosed using more traditional medical screening methods. Furthermore, since such behavioral tests are relatively easy to administer, and are relatively inexpensive they could provide a cost-effective means of screening for harmful effects of toxic chemicals. The goal would be to identify behavioral changes in time to take action to prevent the development of irreversible health impairment. The detection of adverse changes could trigger action to modify the work environment to reduce exposure levels, to make greater use of personal protective equipment for workers, or to move affected workers into jobs where exposure levels are lower or non-existent.

What factors have encouraged the development of the field of Behavioral Toxicology? Interest in this area can be traced back to the latter part of the 1950's. The primary stimulus at this time was the awareness that the Russians had set more stringent standards for air quality than had been set in the U.S. Furthermore, at least according to reports from the Russians, these standards were based at least in part on data obtained from behavioral studies. Although some studies were conducted during the 60's, it was not until the passage of the 1970 legislation that there was significant involvement by behavioral scientists in this country. This involvement was encouraged by research supported by the National Institute for Occupational Safety and Health. Since 1970, the Occupational Safety and Health Act, the awareness that there are over 13,000 toxic substances whose behavioral effects on workers are largely unknown, and the fact that unions are actively supporting research in this area have given rise to the growth of the field.

What are the elements of a Behavioral Toxicology study? How does one go about validating a test battery for diagnosing behavioral effects of exposure? What I would like to do is present one approach that has been used to conduct such a study. Figure 1 presents a flow chart of a typical study.

Insert Figure 1 about here

The study begins as a result of the awareness that a substance is hazardous to workers. This awareness may come from previous research, or from subjective complaints of workers exposed to the substance such as headaches, shortness of breath, weakness, etc. The next step usually involves a review of existing information relating to the chemical substance. This search is designed to uncover information about the known effects on humans or animals, as well as concentration levels required to produce these effects, interactions with other chemicals, and how the chemical is metabolized by the body.

Based primarily on the known effects of the chemical on various human functions, behavioral tests can be selected or developed which are designed to measure these functions. For example, if the chemical is believed to impair central nervous system functioning, a test would be selected that is known to reflect the level of central nervous system functioning, such as critical fusion frequency, or a measure of information processing speed. On the other hand, if the chemical is believed to affect the actions of the peripheral nervous system, it would make sense to select measures that would be sensitive to impaired peripheral functioning, such as a measure of finger dexterity or tremor. In developing this test battery using either previously developed tests or by constructing new measures, it is important to ensure that the battery possesses content validity. In this case content validity refers to the extent to which the tests in the battery adequately sample the range of potential behavioral effects associated with exposure to the chemical.

Once the test battery has been assembled, the next step involves data collection and analysis to investigate its validity. The most rigorous procedure for validating the test battery involves establishing relationships between some measure of exposure level and test scores--this is sometimes called establishing "dose-response" relationships. Measures of exposure are basically of two types: biological or body burden measures taken from the worker and environmental measures providing indications of the actual concentrations in the workplace. The necessity for measures of exposure normally require the behavioral scientist to work with industrial hygienists, analytical chemists, and medical specialists. The need for such a multidisciplinary study team is perhaps one reason behavioral scientists have only recently become involved in this area of research.

A second approach to validating behavioral test batteries has been used when data regarding exact exposure conditions has been lacking. This approach involves defining groups of workers on the basis of exposure history and the presence of observable clinical signs. Using this design a study might identify three groups of workers, those who have been exposed to the chemical and have observable clinical signs, those who have been exposed but have no observable clinical signs, and a control group that has not been exposed at all. Support for the validity of the test battery would be established if the tests were successful in differentiating between these three groups of workers.

The final stage of a toxicology study, as with any research study involves drawing conclusions and making recommendations. The nature of the recommendations would, of course, depend on the nature of the dose-response relationships established and the purposes for which the study was conducted. Types of outcomes that may result from a behavioral toxicology study may include: (1) recognition of the need to lower exposure levels; (2) the need to reduce exposure duration through change of work procedures or job rotation; or (3) the identification of behavioral measures that could be used as part of a routine health screening procedure.

Examples of Test Batteries

Let us now take a look at some test batteries that have been used in some recent behavioral toxicology studies. Figure 2 presents the battery used in a study of the behavioral effects of exposure to Inorganic Mercury (Chaffin, D.B., and Miller, J.M., 1974).

Insert Figure 2 about here

Inorganic mercury is a substance used in the chlor-alkali industry as part of the process used to produce chlorine and caustic. Its primary effects are related to the central nervous system. As can be seen in Figure 2, five of the measures were significantly related to measures of mercury exposure: EMG, tremor, discrete positioning, finger tapping speed, and toe tapping speed.

Figure 3 shows the functions measured by a test battery developed to study the behavioral effects of exposure to lead (Morgan, B.B., and Repko, J.D., 1974).

Insert Figure 3 about here

Lead has a number of uses in industry, one of the principal ones being in the manufacture of storage batteries. The principal toxic effects of lead have been shown to include anemia, colic, and neurological impairment--both central and peripheral. As indicated in Figure 3, four measures were significantly correlated with lead exposure: signal detection; tremor; muscular strength, endurance and recovery; and eye-hand coordination.

Figure 4 lists the components of a test battery used to study the toxic effects of worker exposure to carbon disulfide (Hanninen, 1971). Carbon disulfide is a solvent used in the manufacture of viscose rayon. Although the solvent has been shown to have a wide range of effects, the most consistently noted effect has been peripheral neuropathy--impairment of the peripheral nervous system. This test battery was administered to three groups of workers. Group 1 contained 50 exposed workers who were diagnosed as having clinical carbon disulfide (CS_2) poisoning. Group 2 contained 50 workers who had been exposed for at least five years but had no clinical signs. Group 3 contained 50 unexposed workers who provided a control group.

Insert Figure 4 about here

Table 1 presents selected results from the study. The third column presents significance levels involving differences in test scores between the clinically poisoned and unexposed workers. Seven of the 11 measures included in this table were able to differentiate between these two groups. These differences indicate that behavioral measures are able to differentiate between groups that can also be differentiated on the basis of clinical signs. Although this is an interesting finding, it is not as significant as the results appearing in column two. Column two deals with differences in test scores between two groups that cannot be differentiated on the basis of observable clinical signs. Four of the intelligence measures significantly differentiate these groups at varying levels of significance, as well as three scores derived from the Santa Ana dexterity test. The Santa Ana results seem particularly important since carbon disulfide is known to affect the peripheral nervous system. Peripheral nerve damage would be expected to significantly reduce the speed and accuracy of finger movements as is demonstrated by these results.

Insert Table 1 about here

The lack of detailed exposure data, however, makes this data open to the alternative interpretation that the results were due to differences in sampling rather than differences in exposure. Although this alternative explanation seems more plausible for the differences in intelligence measures than in the dexterity measures, some doubts are cast on the data. From the point of view of a government regulatory agency considering use of the data to modify an existing safety standard that may require companies to spend millions of dollars to modify the workplace, such alternative explanations must be ruled out whenever possible. In addition, this study was conducted in Europe. Before the conclusions can be generalized to a different work environment and a different culture, additional research is needed.

Thus, the Westinghouse Behavioral/Safety Center is conducting what is essentially a replication of the Hanninen study.[2] In order to rule out the alternative explanations,an attempt will be made to relate scores on tests to quantified indices of exposure. The test battery includes those tests found to be most effective in the Hanninen study as well as other measures suggested by a comprehensive literature review. This test battery will include the following:

1. Santa Ana dexterity test (right, left and coordination)
2. Simple and choice reaction time
3. Critical flicker fusion
4. WAIS subtests (digit span, digit symbol, and block design)
5. Neisser Letter Search (vigilance)
6. Feeling tone checklist (subjective fatigue)

This battery will be administered to approximately 150 carbon disulfide exposed workers and 45 unexposed workers in an effort to determine the generality of Hanninen's results with American workers. In addition,the same test battery will also be administered to a small group of workers exposed to perchloroethylene (a dry cleaning solvent) using a repeated measures design to study effects of exposure to this solvent. Data collection is scheduled to begin on this project in a couple of months.

As you can see, research in the field of Behavioral Toxicology is just beginning to uncover some of the effects on workers of exposure to the more than 13,000 toxic chemicals in the workplace. The problems are urgent, the potential rewards and benefits to workers and society are great. As a behavioral scientist, the field truly provides a "New Frontier" for the application of our testing technology. Let me invite you to accept the challenge of this "New Frontier" and actively explore its potential applications for improving conditions of both uniformed and non-uniformed personnel in the military environment.

---

[2] Supported by National Institute for Occupational Safety and Health Contract No. HSM 99-73-35 "Behavioral and Neurological Evaluation of Workers Exposed to Industrial Solvents" Carbon Disulfide; Tetrachloroethylene."

REFERENCES

Chaffin, D.B., & Miller, J.M. Behavioral and neurological evaluation of workers exposed to inorganic mercury. In Xintaras, C., et al. (Eds.) *Behavioral Toxicology*, Washington, D.C.: U.S. Department of Health, Education and Welfare, Publication No. (NIOSH) 74-126, 1974, 214-239.

Hanninen, H. Psychological picture of manifest and latent carbon disulfide poisoning. *British Journal of Industrial Medicine*, 1971, *28*, 374-381.

Morgan, B.B., & Repko, J.D. Evaluation of behavioral functions in workers exposed to lead. In Xintaras, C., et al. (Eds.) *Behavioral Toxicology*, Washington, D.C.: U.S. Department of Health, Education and Welfare, Publication No. (NIOSH) 74-126, 1974, 248-266.

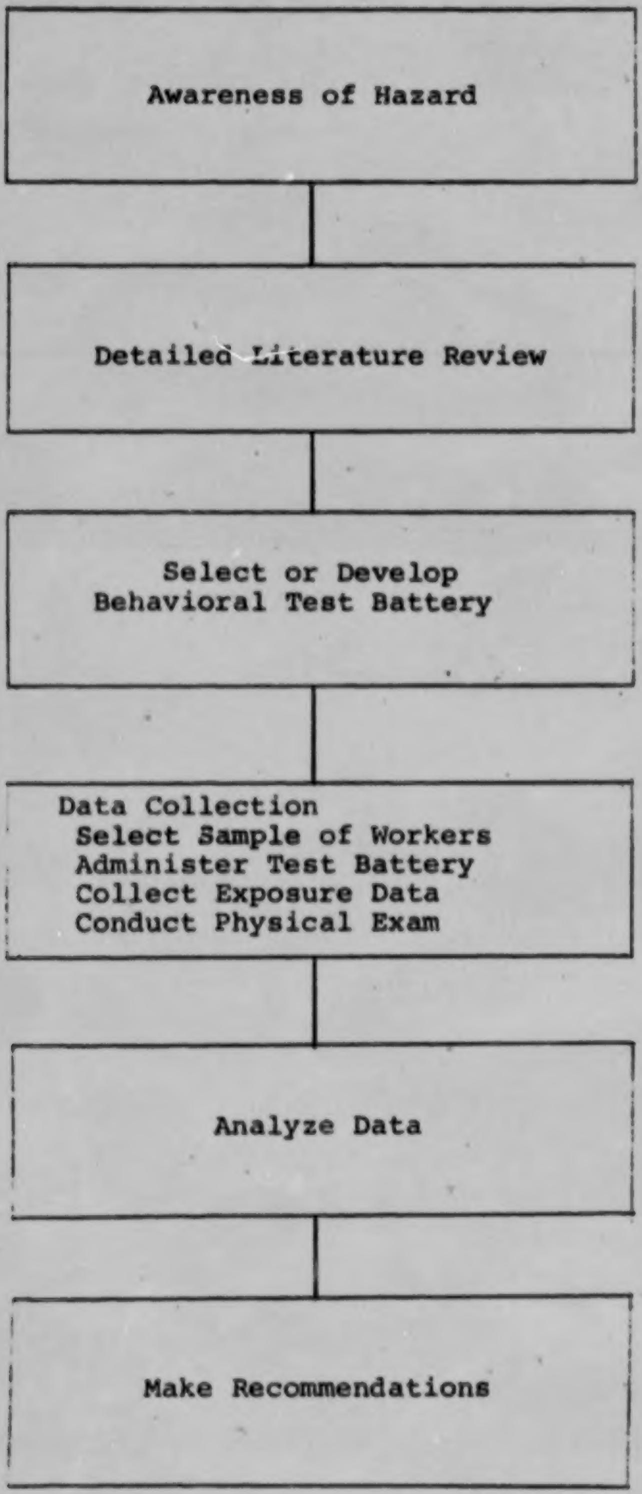

Awareness of Hazard
Detailed Literature Review
Select or Develop Behavioral Test Battery
Data Collection
Select Sample of Workers
Administer Test Battery
Collect Exposure Data
Conduct Physical Exam
Analyze Data
Make Recommendations

Figure 1. Outline of Behavioral Toxicology Study

Surface EMG *
Forearm Tremor *
Localized Muscle Fatigue
Discrete Positioning (Eye-hand coordination) *
Simple Reaction Time
Two-Choice Reaction Time
Four-Choice Reaction Time
Finger Tapping Speed *
Toe Tapping Speed *
Blind Toe Pointing Accuracy
Pencil Flipping Speed

Figure 2. Test Battery for Mercury Study

* Indicates significant correlation with exposure indices (Chaffin, D.B., and Miller, J.M., 1974)

Signal Detection *
Pattern Discrimination
Mental Arithmetic
Visual Acuity
Auditory Acuity
Tremor *
Muscular Strength, Endurance and Recovery *
Eye-hand Coordination *
Immediate Recall *
Subjective Feelings

Figure 3. Test Battery for Lead Study

* Significantly correlated with exposure measures (n = 193) (Morgan, B.B., and Repko, J.D., 1974)

- Intelligence test
    - Digit Span (D Sp)
    - Similarities (Sim)
    - Block Design (BD)
    - Picture Completion (PC)
    - Digit Symbol (D Sy)

- Other abilities test
    - Bourdon Wiersma speed (BW sp): number of rows
    - Bourdon Wiersma errors (BW err): number of errors
    - Bourdon Wiersma fluctuation (BW fl): time variation in completing a row
    - Santa Ana (SA): right hand performance
    - Symmetry drawing (Sym): number of reversions
    - Benton (Ben): number of correct figures

- Personality
    - Rorschach: adaptability (Ro Ada)
    - Rorschach, emotionality (Ro Emo)
    - Rorschach, spontaneity (Ro Spon)
    - Rorschach, rational self control (Ro Rat)
    - Rorschach, original intellectual activity (Ro Orig)

- Psychomotor behavior
    - Mira I: length of lines (Mi I)
    - Mira II: distance of lines (Mi II)
    - Mira III: size of angles (Mi III)
    - Mira IV: left-hand deviation (Mi IV)
    - Mira V: right-hand deviation (Mi V)
    - Mira VI: form level of straight lines (Mi VI)
    - Mira VII: form level of broken lines (MI VII)
    - Mira VIII: performance time (Mi VIII)

- Subjective symptoms
    - Symptoms of neurovegetative lability (Labil)
    - Sleep disturbances (Sleep)
    - Symptoms of fatigue (Fatig)
    - Symptoms of neurotic behavior (Neur)

Figure 4. Test Battery for Carbon Disulfide Study
(Hanninen, 1971)

Table 1. Selected Results from Carbon Disulfide Study (from Hanninen, 1971)

Test	Significance of differences(t-tests)		
	Poisoned vs. Exposed	Exposed vs. Control	Poisoned vs. Control
Intelligence			
1. Digit Span		+	
2. Similarities			
3. Picture Completion		++	++
4. Block Design		++	
5. Digit Symbol	++	+	+++
Other Abilities			
6. BW speed	+++		+++
7. BW errors			
8. BW fluctuation	++		+++
9. SA right	+++	+++	+++
10. SA left	+++	+++	+++
11. SA coordination		+++	+++

+ $P < .05$

++ $P < .01$

+++ $P < .001$

ENTRANCE SCORES AND TRAINING PERFORMANCE FOR TWO GROUPS OF COAST GUARD SCHOOL INPUT

Anita S. West
Ralph E. Williams

Denver Research Institute

Abstract

The present Navy Test Battery standards for entrance into the Coast Guard Class A schools evolved from the Navy's cutting score standards for comparable training. These levels were influenced over the years by a very favorable Coast Guard selection climate, attributed at least partly to the military draft. During this time, Coast Guard volunteers were generally above average mentally and often college educated. As part of a study to review both the curricula of selected Class A schools and the predictive ability of these standards, background and school data were collected (from fiscal 1973) on over 1,700 trainees. The experiences of students coming directly from recruit training were examined separately from those coming from field billets. Although there are differences from school to school, boot input generally have slightly higher entrance scores and field input have slightly higher final scores. Some of the hypotheses to explain these phenomena are discussed. Training for recruits has traditionally been provided as soon as possible after procurement, since most personnel were not expected to reenlist. Anticipated higher retention rates and various enlistment procedure modifications may affect the optimum ratio of field to boot input if the variables affecting success can be identified.

I. Background

Selection for training in the Coast Guard Class A schools has most often occurred in one of two different ways: either directly upon graduation from recruit training (boot input) or after having served in a nonrated operational billet as a Seaman (field input). In general, persons with demonstrated ability to succeed in schools of their choice, as evidenced by performance on the Navy Basic Test Battery, were admitted into the schools upon the successful completion of boot camp. However, the adequacy of the standardized test batteries in screening school candidates has always been questioned by some. The traditional procedure for providing equitable opportunities for schooling has been to grant motivated persons who have exhibited the potential to become petty officers another opportunity to qualify for training. Most often this procedure has been to merely reduce cutting score requirements (5-10 point waivers) for recommended field applicants. The resulting proportion of field to boot school input was approximately 40 to 60 percent during the period immediately preceding initiation of the all volunteer force.

There are several factors operating at present that are having an effect upon this ratio. These factors and the directions in which they appear to be influencing the percentage of field input are shown below.

(1) Changes in qualifications as measured by test battery scores. (+)

Although the Coast Guard is and has been staffed with volunteers, it has enjoyed a favorable selection situation in the past partly because of the many qualified enlistees who chose to satisfy their military obligations by enlisting in the Coast Guard. School cutting scores, once adopted from U.S. Navy standards for comparable training, were changed over the years to reflect this favorable enlistment picture. During this period, it was considerably more efficient to generate standards more likely to err in the direction of rejecting a potentially successful candidate than in accepting a potential failure. The initiation of the all volunteer military has affected the number and therefore the qualifications of new accessions. Accepting more recruits with reduced qualifications into the service tends to increase the number of candidates who do not qualify for Class A schools from recruit training.

(2) Reduction in continued enlistment growth. (-)

Previous training quotas presumed continued enlisted growth, which for the past two years has not materialized. Consequently, there has been some overproduction of Class A school graduates that has reduced school openings and subsequently resulted in cutting back on field input.

(3) Sensitivity to the need for individual opportunity. (+)

There appears to be more sensitivity to the need to grant individuals opportunities for receiving schooling and becoming rated than ever before. The efficient, but occasionally discriminatory, probabilistic approach used in the past in private employment and by institutions of higher education as well as by the services is being rejected as unacceptable by those who see the "likelihood to fail" label as perhaps valid in the aggregate, but unfair to the individual candidate who despite a low success profile has unidentified compensatory abilities that will allow him to succeed.

(4) Sensitivity to unrecognized biases in testing instruments. (+)

Similar to the reasoning involved in number (3) above, it has become increasingly apparent that most standardized tests discriminate in favor of experienced test-takers and against those whose cultural backgrounds digress from the norm. This recognition leads also to an attempt to increase opportunities for those who fail to meet present standards based wholly upon test battery scores. Also in this regard new screening devices are being recommended and developed.

(5) Guaranteed school program. (-)

The guaranteed school enlistment program guarantees the recruit entry into a school of his choice upon the successful completion of recruit training. This program, viewed as a necessary incentive for recruiting sufficient enlistments, nevertheless severely restricts opportunities for those who do not take advantage of the guaranteed school enlistment option. The initiation of this program, concurrently with the effects of limited growth, curtails training opportunities for field personnel (and also for regular enlistment boot input).

Although these factors operate to affect the field/boot ratio in both directions, numbers (2) and (5) seem to have had the greater impact, and input from the field has been sharply reduced. The proportions have changed from 40:60 to approximately 25:75 (field:boot).

II. Source of Data

The University of Denver Research Institute is currently performing a research study for the U.S. Coast Guard to examine both the cutting scores and curricula of selected Class A schools. Cutting scores are being studied to evaluate their predictive ability and curricula are under examination for their relevance to actual on-the-job requirements of graduates at the E-4 level. The DRI investigation has reviewed test battery scores, educational attainment variables, and school performance data collected from fiscal 1973 on over 1,700 trainees for this study. Over 1,100 of these people were identifiable as having entered the schools from either boot camp or field assignments. In performing the analysis of the predictive ability of current cutting scores, correlations between final score and various test battery scores were examined. From this analysis new recommeded cutting scores were generated and regression equations were developed in order to examine post hoc the relative efficiency of the two sets of standards. During these data processing manipulations, it began to appear as though (for the 1973 data) field and boot inputs actually represented two different populations. Although there were not significant differences in the test battery scores or in the final class performance scores in all of the schools, there were significant differences between the groups on both measures in the total sample, and covariance techniques showed that when differences in entry scores were accounted for, field personnel performed significantly better than boot personnel in almost all cases. Since current field input are often as qualified or even more qualified than boot input (because of the effects of limited Class A opportunities and the guaranteed school enlistment program), the differences between boot and field populations are apt to become obscured in fiscal 1974 and 1975 data and it was considered important therefore to report these differences in the 1973 data.

The policy of sending boots directly to Class A schools is based upon the assumption that draft motivated personnel would not reenlist, and therefore for the service to benefit most from their training it should be provided as soon as possible after procurement. With the changeover to the all volunteer force, the assumption is that enlistees will be coming into the Coast Guard as a career choice decision, and the expectation is for higher retention rates. Therefore, the decision to send all qualified personnel to school immediately following recruit training may not be in the best interests of the service. Recommendations for new selection procedures can include field and boot parameters if the variables affecting success in each school, as well as on the job, can be sufficiently identified. This paper is a first attempt to isolate parameters of success differentially for field and boot input. Because these data were not generated for this purpose, most of the hypothetical factors involved can only be recommended for future examination.

III. Summary Presentation of Data by Field and Boot

Data from seven of the Class A schools for which trainee information was made available for the DRI study identified separately those persons entering the training schools from the field and from boot camp. Examination of GCT scores, educational attainment and final class scores for these seven schools showed that the boot group performed significantly higher on the test battery (at the .01 probability level) and that the two groups were not different in educational attainment. See Table 1 below.

Table 1

Comparison of Field and Boot on Three Measures
(Fiscal 1973 - 7 Schools)

	GCT*			Education (Highest Grade Attained)			Final Scores**		
	N	$\bar{X}$	σ	N	$\bar{X}$	σ	N	$\bar{X}$	σ
Field	417	57.19	7.51	418	11.99	1.12	416	812.91	59.06
Boot	696	60.48	6.70	697	12.00	1.11	696	803.99	62.05

*t significant $p > .01$
**t significant $p > .05$

Also shown in this comparison are the final mean scores for field and boot across all seven schools. (Final score in most instances is an average of the weekly scores reported by each school for each trainee. Although there were no significant differences between field and boot on attrition rates, academic disenrollees were assigned a final score of 600.) Final scores are significantly different at only the .05 probability level. However, controlling for the difference in GCT scores by covariance, the final scores are significantly different at the 0.001 probability level.

Covariance provides the investigator with a measure of control over individual differences. Analysis of covariance first adjusts the criterion (final score) based on the covariate (GCT score) by means of regression and then examines the final test scores for differences as if the two groups had had the same GCT scores. This is an important calculation if future field and boot inputs are apt to be very much alike on entrance test scores.

The results from Table 1 appeared interesting enough to warrant analyzing data from each school by field and boot groupings. The results of this analysis are shown in Table 2, where the levels of significance vary considerably and in fact show no differences in either entry scores or performance for some schools

GCT scores.

Although the means for field GCT scores are lower than boot in all but one group (RM, with only 23.6 percent field input), the differences are significant only for the SK and MK school at $p > .01$. GCT scores for the EM school are significantly different, but only at the .1 probability level.

Final scores.

The means for final scores are significantly higher for the field group for the DC, the SK, RM, and MK schools at $p > .01$. Covarying the GCT scores again we see significance in the difference in final scores for the EM school at the $p > .05$ level and even for the RM school (where GCT's were slightly higher for the field group) at the $p > .1$ level. The two schools that show no significant differences are the GM and RD schools. In no cases were field GCT's significantly higher than boot nor were field final scores significantly lower than boot.

TABLE 2

GCT, EDUCATION, AND FINAL SCORE MEANS FOR SEVEN SCHOOLS

School		GCT		Education		Final	
	N	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ
EM							
Field	62	45.40	8.53	11.83	1.03	814.37	102.61
Boot	78	56.32	7.01	11.93	.93	816.75	34.82
		F>B significant at p. >.1				*Co-varying GCT F Sig. p. >.05*	
GM							
Field	24	57.00	6.84	11.60	1.07	780.36	97.91
Boot	18	58.58	5.81	11.58	.67	808.83	81.94
DC							
Field	95	57.30	6.71	12.28	1.86	782.31	39.45
Boot	118	58.12	5.72	12.01	1.01	763.61	56.24
						F>B p. >.01	
RD							
Field	20	57.80	5.84	12.10	1.00	800.60	61.74
Boot	54	59.25	7.44	12.00	.97	808.49	67.20
SK							
Field	101	56.55	8.32	12.25	1.17	850.08	71.71
Boot	143	60.03	6.68	12.54	1.54	844.67	80.39
		F>B t significant at p >.01				*F>B p >.01*	
RM							
Field	38	59.76	6.28	11.68	2.02	832.61	57.27
Boot	161	57.55	6.97	11.98	1.13	800.70	59.02
						F>B p >.01 co-varying GCT sign. p >.1	
MK							
Field	77	54.00	7.29	11.88	1.17	801.49	56.96
Boot	124	56.67	8.11	11.65	1.25	790.18	57.33
		F>B p >.01				*F>B p >.01*	

IV. Summary and Conclusions

Our data show that the field personnel were generally less well qualified than boot personnel according to test battery scores, and yet their school performance scores were generally higher. When GCT scores were covaried, analysis of variance showed that the significance levels for the differencs in final scores rose sharply. With the expectation that field people will become more and more similar to boot personnel on entry score variables, it becomes important to attempt to become more knowledgeable about the beneficial effects of assignment to an operational billet prior to Class A schooling.

The schools for which the phenomenon of improved performance among field personnel was most prominent were the Storekeeper and Machine Technician, followed by the Damage Control and Radioman schools. The schools in which field inputs did not perform as well as boots were Gunners Mate and Radar Technician. In the case of the storekeeper rating it can be postulated that exposure of all seamen to requisition forms and procedures has an effect upon the expectations of what is required in the rating and therefore success in training. In this regard it would have been valuable to have had information on the particular billet from which each of the trainees was assigned. Because of the need for preliminary safety training, it may be that field input into the Gunners Mate school (for which boot out-performed field) had not been serving in a GM-related billet prior to training. The degree to which exposure to the Coast Guard in general and to particular ratings influences success in school can be estimated from current experience if these data are recorded and collected.

Also unavailable in the data examined was the variety of personal background variables that may also have affected success differentially for the field and boot people. The study of parameters such as age, sex, prior employment may allow counselors to suggest in the future, for example, that a young man between the ages of 17 and 19 who shows a preference for RM training, has never been employed, and who has diagnostic scores in a given range would benefit from serving for short times in both technical and nontechnical billets prior to his 20 some weeks of training. The reliability of rating selection and the increased success with schooling may offset the disadvantages of postponed training. For others, prompt schooling and placement may be indicated. A third alternative could be the nonpenalizing option for immediate or postponed schooling. If we are thinking of a smaller but more career-oriented service, then the additional time and resources allocated to training and job assignment procedures become increasingly more efficient and have the advantage of being both more equitable and more satisfying for most people. Further, it should reduce many problems for both the training personnel at the schools and the receiving billet officer to whom the new graduate is assigned.

Acknowledgements

The results reported herein were obtained from work performed under Contract No. DOT-CG-40353-A from the U.S. Coast Guard. The authors wish to acknowledge their appreciation to the technical contract monitor, Mr. Joseph Cowan, for his contributions to this effort. Also acknowledged with appreciation are the data processing efforts supervised by Sandra Bleistein, DRI System Analyst.

Selection of Officer/Students for Graduate Education[1]

R. S. Elster
Naval Postgraduate School

W. H. Githens
Navy Personnel Research
and Development Center

The U.S. Navy and the other U.S. services spend millions of dollars every year on the graduate education of their officers. In recent years, about 570 U.S. Navy officers a year have received Navy funded graduate degrees at the Naval Postgraduate School (Congressional Record, 1973, p. 617). The justification for these educational programs has been that certain Navy officer billets require the incumbent to have such an education to perform optimally in the job (SECNAVINST 1520.4A). The best payoff to the Navy comes, then, when officers are selected for educational curricula in which they perform well, and in which they develop skills they wish to apply in their later Navy billets.

Present Selection Procedures

The Navy's Bureau of Personnel (BuPers) derives the educational quotas for the different curricula by juxtaposing billets requiring a graduate officer with the populations of officers having that education. These educational quotas are then reviewed in light of existing monetary and manpower constraints. Ultimately, shortfalls are published; classified by educational area, year group and designator (e.g., aviator). BuPers then provides the shortfall data to the Postgraduate Selection Board. The Postgraduate Selection Board is annually formed by BuPers and requires a month or more to complete its work. The Board's members are ranking officers from each Navy designator, and many of them hold graduate degrees. The Board's task is to examine the records of officers who could be selected for graduate education. These records include the officer's fitness reports and his prior academic records. Occasionally, the officer's scores on the Graduate Record Exam (GRE) would also be in his records. Also available to the Postgraduate Selection Board for most officers are their curricular preferences. Given the officers' records and preferences, and the educational quotas, the Postgraduate Selection Board proceeds with the selection of Navy officers for graduate education programs. The FY 75 Board considered 15,406 officers and selected only 1,285. In making their evaluation

---

[1] This research was sponsored by the Navy Personnel Research and Development Center, San Diego, CA.

of each of the thousands of officers, prior academic work is supposed to be weighted 40% and Navy performance (as assessed by fitness reports) 60%. Once the Postgraduate Selection Board has finished its month or more of work, the list of officers' names is sent to the officer detailing section of BuPers. The detailers have the final responsibility for assigning officers to graduate curricula.

Method

The investigators felt that the methods and procedures for selecting Navy officers could be made less time and manpower consuming, and, probably, more valid. Consequently, a concurrent validation study was designed and conducted, using as subjects U.S. Navy officer/students on-board at NPS.

The Strong Vocational Interest Blank (SVIB), the GRE and a biographic data collection form were administered to the groups shown in Table I during fiscal year 1974.

Table I

Curricula and Numbers of Students

Curriculum	Number of Students
Ops. Res./Sys. Anal.	80
Computer Systems Mgt.	48
Computer Science	24
Aero Eng.	103
Communications Mgt.	42
Sys. Acq. Management	52
Management	130
TOTAL	479

NPS grade averages (TQPRs) were also gathered from these officer/students.

Preliminary Analyses

As the grade average was to be used as the criterion measure in this research, it was necessary to understand its behavior. The first question addressed concerned the accuracy of self-reported TQPRs.

The NPS academic transcripts of 55 Management students were gathered and the computed TQPRs compared with the self-reported NPS

TQPRs. The correlation between the reported and actual TQPRs was .99, and the means were 3.37 and 3.34 (on a scale where A = 4.00), respectively (Welch, 1974, p. 19). A similar analysis was done with the records of the 42 U.S.N. students in Communications Management. In all 42 cases, the actual and self-reported NPS TQPRs were identical. Obviously, the investigators could use self-reported NPS TQPR as criterion data and avoid retrieving each officer's transcript.

A second study of TQPRs investigated the longitudinal stability of individuals' TQPRs. This investigation was important for two reasons. First, if grades are quite unstable over time, then they might be unpredictable, or different variables might be required to predict TQPRs in different quarters of an academic curriculum. The second reason for doing the grade stability study arose from the relatively small sample sizes in the different curricula. The set of students in a particular curriculum consisted of two or more subsets of students that had been in the curriculum different lengths of time. If TQPRs were stable over time, then TQPR data from students who had finished only a few quarters could be considered as indicative of what their TQPR would be upon completing their curriculum. In other words, if TQPRs were quite stable, the data from the different student subsets in a curriculum could possibly be aggregated. Table II is representative of the patterns of correlations found in the curricula investigated. The correlations are high--even given that they are part-whole correlations. In summary, the correlations indicate that TQPRs after two, or at most, three quarters are very predictive of TQPRs after six quarters.

Table II

Correlations Among TQPRs After Completion of
Different Numbers of Quarters of the Ops. Res./Sys. Anal. Curriculum:
Longitudinal Data (N = 25 U.S.N. Officers)[1]

		Quarters Completed					
		1	2	3	4	5	6
	1	1.00	.95	.92	.92	.90	.89
	2		1.00	.98	.97	.96	.95
Quarters Completed	3			1.00	.99	.98	.97
	4				1.00	.99	.99
	5					1.00	1.00
	6						1.00

[1] Source: Welch, 1974, p. 21

Further analyses of the TQPR data uncovered two other facts: First, the variance among inputs (an input being a group of students entering the curriculum at the same time) on TQPR is much smaller than the TQPR variance within an input in a curriculum; second, an input group's average TQPR tends to sneak upwards as their time spent in the curriculum increases.

Because of the results of the analyses conducted on NPS TQPRs, the decision was made to standard score the TQPRs for each input into each curriculum. Thus, the standardization process yielded equal TQPR means and standard deviations for all of the inputs and curricula.

TQPR Prediction Study Results

Data from the Graduate Record Exam (GRE), Strong Vocational Interest Blank (SVIB), and a biographic data form were used in attempts to predict TQPRs in four separate curricula: Ops. Res./Sys. Anal., Aero Eng., Communications Mgt., and Management. Multiple linear regression was the primary analytical technique used in the prediction studies. Within each curriculum, developmental and cross-validation groups were randomly formed. Prediction equations generated using the data from a developmental group, and yielding a statistically significant multiple correlation coefficient in the relevant cross-validation group, will be reported in the following paragraphs.

Aeronautical Engineering

Several useful prediction equations were developed for this curriculum; the equation presented here for predicting standardized TQPR is the most face valid one and yielded a multiple correlation of .43 in the cross-validation group of 30 students (Sofge, 1974, p. 52).

$$\begin{aligned} \mathrm{TQPRZ} = &- 4.20 - .635(\mathrm{VAR001}) + .691(\mathrm{VAR004}) \\ &+ .00184(\mathrm{GREV}) + .00322(\mathrm{GREQ}) \\ &+ .00568(\mathrm{INDEX}) - 1.02(\mathrm{BQPR}) \end{aligned}$$

Where: TQPRZ = Standardized TQPR
VAR001 = Naval Academy Graduate: yes = 1; no = 0
VAR004 = Lt. or below: yes = 1; no = 0
GREV = GRE Verbal Score
GREQ = GRE Quantitative Score
BQPR = Self-reported baccalaureate grade average
INDEX = BQPR × Qual (Qual is an estimate of the quality of the individual's baccalaureate institution from Hoyt's data.)

Management

A rather simple prediction equation was found to be useful in predicting TQPRs in the Management curriculum.

$$\text{TQPRZ} = -3.94 + .0059(\text{GREQ}) + .6897(\text{VAR001})$$

Where: TQPRZ = Standardized TQPR
GREQ = GRE Quantitative Score
VAR001 = 1 if chronological age $\leq$ 30.
VAR001 = 0 if chronological age > 30.

This equation yielded a multiple correlation of .65 in the cross-validation sample (N = 20).

Operations Research/Systems Analysis

The multiple regression analysis conducted with the data from this curriculum yielded an equation having in it three predictor variables.

$$\text{TQPRZ} = -2.3445 + .00322(\text{INDEX}) + .4067(\text{VAR001}) + .2668(\text{VAR002})$$

Where: TQPRZ = Standardized TQPR
VAR001 = 1 if the officer speaks a foreign language
VAR001 = 0 if the officer does not speak a foreign language
VAR002 = 1 if the officer's age is $\leq$ 30.
VAR002 = 0 if the officer's age is > 30.
INDEX = BQPR × Qual (BQPR was self-reported baccalaureate grade average. Qual is an estimate of the quality of the officer's baccalaureate institution from Hoyt's data.)

This equation yielded a multiple correlation of .45 in the cross-validation sample (N = 20).

Communications Management

A multiple regression equation using four predictors was developed for this curriculum (Cook, 1974).

$$\text{TQPRZ} = -\,3.632 + 1.746(\text{4GPA}) + 1.454(\text{VAR001}) - .6546(\text{VAR002}) - .606(\text{3GPA})$$

Where: TQPRZ = Standardized TQPR
4GPA = Calculated grade average in the individual's fourth undergraduate year
VAR001 = 1 if the individual had an undergraduate major in engineering
VAR001 = 0 if the individual was not an undergraduate engineering major
VAR002 = 1 if the officer had graduated from the Naval Academy
VAR002 = 0 if the officer had not graduated from the Naval Academy
3GPA = The individual's grade average in his third undergraduate year

The equation yielded a multiple correlation coefficient of .63 in the cross-validation group (N = 12).

Discussion

The results shown above clearly indicate that academic aptitude and biographic data could be used to select U.S. Navy officers for attendance at the Naval Postgraduate School. Other analyses, not reported here, have shown the SVIB to be useful also in predicting TQPR. The results reported here seem impressive when it is considered that the students in the study had already been through the Postgraduate Selection Board screening--a process which should have led to restriction of range on a number of variables, with a consequent lowering of the multiple correlation coefficients.

For the Navy's Bureau of Personnel to use the findings reported here, biographic data and GRE scores would have to be available for the officers being considered for graduate education. The costs of developing these data would seem to be small when compared to: (1) the time saving that could accrue by streamlining the Postgraduate selection process, and (2) the more effective use of the officer/student human resource that could be achieved through more valid selection methods.

References

1. Cook, J. L., Predictors of Academic Success of Graduate Students in the Communications Management Curriculum of the U.S. Naval Postgraduate School, M.S. Thesis (1974), U.S. Naval Postgraduate School, Monterey, CA

2. Hoyt, D. P., Forecasting Academic Success in Specific Colleges, 1968, ACT Research Report No. 27, American College Testing Program, Iowa City.

3. SECNAVINST 1520.4A, Policy on Graduate Education, U.S. Dept. of the Navy, 2 July 1971.

4. Sofge, C. T., Prediction of Performance and Satistaction of Aeronautical Engineering Students at the Naval Postgraduate School, M.S. Thesis (1974), Naval Postgraduate School, Monterey, CA

5. Welch, T. A., Predictors of Grades and Satisfaction Using the Strong Vocational Interest Blank, M.S. Thesis (1974), Naval Postgraduate School, Monterey, CA

SECTION VI

PERFORMANCE MEASUREMENT AND STATISTICAL METHODOLOGIES

October 1974

AUTOMATED PERFORMANCE MEASUREMENT USED IN QUALITY CONTROL OF THE U.S. COAST GUARD AVIATOR TRAINING SYSTEM

BY

H. Kingsley Povenmire

Department of Transportation
U.S. Coast Guard
AVIATION TRAINING CENTER
Mobile, Alabama 36608

The opinions expressed herein are the private ones of the writer and are not to be construed as official or reflecting the views of the Commandant or the Coast Guard at large.

H. Kingsley Povenmire

H. Kingsley Povenmire
October 1974

Automated Performance Measurement Used in Quality Control of the U.S. Coast Guard Aviator Training System

H. KINGSLEY POVENMIRE, Chief, Training Analysis Support Branch,
Training Division, U.S. Coast Guard Aviation Training Center,
Mobile, Alabama

ABSTRACT

Historically measurement of pilot skills has been difficult to standardize due to the subjective nature of instructor or checkpilot grades and the complexity of the pilot's tasks. Because checkpilot judgment must be allowed to enter into pass-fail decisions on individual aviators it is difficult to analyze the effectiveness of a training program based on instructor or checkpilot scores. Limited objective measurement of pilot performance is provided by automated checkrides in the U.S. Coast Guard Variable Cockpit Training System (VCTS). Tracking skills and some procedural tasks are scored in the context of an operational instrument flight in a groundbased helicopter flight simulator. Performance scores consisting of frequency and time out of tolerance of preset limits on twelve parameters are collected for each segment of eight maneuvers. Comparisons of these scores show discrimination between groups of pilots with different experience levels.

BACKGROUND

For many years industry has used statistical techniques in quality control of production and marketing. In addition to well known sales graphs and production quotas, records are kept which assist management in determining cost effectiveness of suggested changes in the production process. Statistical analysis may also provide one of the first indications that product quality is deteriorating.

Similar techniques might be used in aviation training to assist in making decisions concerning modification of the training process, or to assure that there is no deterioration in the training system. To accomplish this we must have accurate objective data concerning the quality of the training product.

PROBLEM

Evaluating a student's skill at flying an aircraft is quite complex. For years instructors have graded daily flights using some sort of subjective grading scale. This type of device is widely used in military flight training today. Students receive a single mark for each maneuver based on a composite of flight proficiency, judgment, and some intangible evaluation of their confidence level.

This system has some obvious statistical disadvantages. It is difficult to determine whether a given checkpilot is predisposed toward higher scores, or actually observing better performances. In addition there are factors which may subconsciously affect day to day changes in the criteria used by an individual checkpilot.

Attempts to provide checkpilots with objective scales for grading pilot skills have measurably increased the reliability between observers, however some variance still exists (Smith, et al, 1952; Caro, 1968; Povenmire, et al, 1970; and Koonce, 1974).

Automated recordings from flight instruments provide an abstract description of pilot performance which completely eliminates the possibility of check pilot bias (Jensen, et al, 1971 and Knoop, 1973).

As important as it is to have objective performance data, it must be recognized that complete reliance on data derived from instrument readings decreases the scope of evaluation. One of the most important aspects of pilot skill involves judgment and decision-making strategy which cannot be evaluated from instrument indications. As pilots gain in experience and confidence they become better able to determine alternative courses of action, and more likely to make legitimate deviations from standardized procedures. Unfortunately, objective scoring of individual judgment is presently not possible. We must necessarily rely on subjective grades and comments by checkpilots for evaluation of students' confidence and judgment. The checkpilot must ultimately make a pass-fail decision considering all facets of a pilot's performance.

PROCEDURE

Training System

The Variable Cockpit Training System (VCTS) located at the U.S. Coast Guard Aviation Training Center in Mobile, Alabama utilizes a highly sophisticated flight simulator for the HH-52A and HH-3F helicopters now in service throughout the Coast Guard. This simulation complex was built by Reflectone, Incorporated, utilizing a Datacraft 6024/3 computer with a high speed disk operating system. Each cockpit is installed on a six-post motion system providing all motion cues experienced by helicopter pilots except sustained G forces. In addition to the simulators themselves, the VCTS incorporates a well-organized training system developed in cooperation with the Human Resources Research Organization (HumRRO) who, along with the Naval Training Equipment Center, assisted in the design and procurement of the simulation complex (Hall, et al, 1969 and Caro, et al, 1969). The entire project was under the direction of CDR Gilbert E. BROWN who then became the first Chief of the Synthetic Training Branch.

Many advanced training capabilities are programmed into the simulators including performance playback, automated demonstrations of selected maneuvers, and automated performance scoring. Two students from different Coast Guard air stations are teamed as pilot and copilot alternating throughout the course. The instructor is seated behind them in a position to observe as well as operate the simulator. All simulation controls are available to the instructor. He may insert or delete malfunctions, freeze the simulator to discuss problems, replay student performance to point out missed cues, or cause the simulator to demonstrate a troublesome maneuver. Training flights are conducted in an operational context to add realism to the decision-making process. The student is required to consider geographic position, and mission requirements in responding to various emergencies.

The VCTS is used in three different courses. The Qualification Course, given only in the smaller HH-52A, provides rotary wing qualification for fixed wing pilots. Students are trained to proficiency in basic flight procedures including selected emergencies on instrument flights in the simulator. They then go to the aircraft with the same instructor to learn visual maneuvers such as hovering and landing which require visual simulation not presently available. Simulator flights are interspersed with this phase to complete the systems management and emergency training. Students are then assigned to a variety of instructors for the operational maneuvers involved in search and rescue. This course usually lasts five to six weeks depending on the proficiency and learning rate of the student.

Rotary wing pilots transitioning from one helicopter to another go through the Transition Course in the appropriate aircraft. VCTS and aircraft flights are interspersed as in the Qualification Course except that less time generally is required for students to learn basic helicopter control. Students usually complete this course in four to five weeks.

Each pilot in the Coast Guard must annually demonstrate instrument flying proficiency in order to obtain his instrument renewal. Since the implementation of the VCTS all Coast Guard helicopter pilots come to the Aviation Training Center for the Proficiency Course which fulfills this requirement. This course, which lasts approximately one week, is taught entirely in the simulator, terminating with an automatically scored instrument checkride. The Coast Guard is the only military service to date that has eliminated the requirement to perform this annual instrument checkride in the aircraft. Our checkpilots feel that simulated checkrides are more comprehensive and indicative of pilot ability than those previously performed in the aircraft.

Quality Control

One vital feature of the training system is the provision for continuous evaluation and revision. A Quality Control group made up of instructors from within the Training Division is responsible for the development and implementation of this system. Quality control in aviation training is intended to serve two functions. First, it should provide feedback regarding changes in the quality of instruction. Second, it should be responsive to new operational procedures requiring changes to the course objectives.

When this quality control system is fully implemented it will receive inputs from several sources. Written critiques are required from each student at the completion of his course. These have been very responsive, reflecting a high level of thought and effort on the part of the students. Many course changes have resulted from comments on these critiques. Questionaires will be sent to each unit who receives a newly designated graduate of the Transition or Qualification Course. Approximately two months after he reports, the unit will be asked to provide feedback of the effectiveness of training as related to the operational task. This effort should provide information on additional training that aviators receive at various units. Training Division instructors are in an excellent position to suggest changes because of their involvement in the total training program.

In addition to these narrative inputs, training time required by each student to complete his respective course is recorded as well as a breakdown of his flight experience.

Objective performance data from the automated checkride described below provides a touch stone for long-term evaluation of overall training quality.

Automated Checkride

Although it is recognized that performance scores based on recordings of aircraft performance are only partially descriptive of pilot ability, there is a positive correlation between these objective performance scores and subjective checkpilot evaluation (Knoop, 1973 and Koonce, 1974). The use of objective automated checkride scores may be justified in comparing groups of students, but not as the sole basis for grading individuals.

The automated checkride incorporated in the VCTS records frequency and time out of tolerance for twelve flight parameters. A simple computer language called SANSKRIT is supplied by Reflectone to allow Training Division personnel to design checkrides.

Each maneuver is broken down into segments for scoring purposes. Errors are accumulated and stored for as many as nine segments of a maneuver. Each segment may be further divided into as many as 15 blocks to allow changes in scoring within the segment, or to ensure proper sequencing through the segment. The checkride is automatically sequenced when the student reaches the "END CRITERIA" of each block. Figure 3 shows an example of the programming of an instrument climb which allows for changing airspeed requirements as the student levels off at his assigned altitude of 3000 feet. Maneuvers are ordered in such a way as to simulate an operational instrument mission. The instructor acts as an FAA controller giving clearances and weather information which would be heard on an actual flight. He also records information which will not show up on the recorded checkride such as cockpit procedures, planning ahead, or how well the student coordinates with his copilot. Prior to the checkflight the student plans the flight as he would in the real world including all required paperwork. A typical checkride commences with an instrument takeoff from Bates Field in Mobile destined for Gulfport 30 miles away. The student is graded on his takeoff, his airways tracking as he flies to Gulfport, and his holding procedures simulating delays for traffic or weather. He will make several approaches down to Minimum Descent Altitude. Several malfunctions are automatically inserted along the way to allow the instructor to subjectively grade his judgment and troubleshooting strategy.

Establishing Criteria

Special attention must be payed to the designing end criteria for each block in order to complete as many checkrides as possible. End criteria must be met not only by any of the acceptable procedures, but by any conceivable procedure.

Many operational instrument maneuvers allow alternative procedures which are equally correct. Scoring criteria must be developed to allow any of the acceptable alternatives to be scored equally. This imposes limits on the number of items that can be scored during a given block. For example, if airspeed in the holding pattern is left to the discretion of the pilot, it cannot be scored.

The computer automatically prints out a record of error scores for each maneuver at the end of the checkride provided that all end criteria have been met. The raw data for as many as six checkrides are stored on the main trainer disk. These records are removed as they accumulate and transferred for permanent storage on a separate disk containing programs for complex data analysis and reporting.

RESULTS

Although the number of checkride performance records is not yet large enough to provide stable samples, we have been able to make a preliminary evaluation of the scoring system by comparing 23 HH-52A Proficiency students who rarely fly instruments during the year with nine instructors from the HH-52A Training Branch who are extremely familiar with the simulator, the instrument environment, and the scoring criteria due to their responsibilities for standardization of Coast Guard pilots. These comparisons show clear discrimination between the two groups. Figure 4 shows the distribution of each group based on total time out of tolerance for the checkride. When these scores are broken down into the 64 individual parameters scored within the seven maneuvers the strength of the comparison is upheld. Of the 58 parameter scores which favored the instructors, 38 were significant ($\underline{p}<.01$). Of the six scores where the order was reversed, no significant differences occurred. Total scores for each maneuver are shown in Table 1, showing the increasing discrimination that occurs as these scores are combined.

At present the limited number of recorded checkrides available does not allow significant comparison between groups within the Proficiency Course, much less within either of the other courses. Our small yearly quota of approximately 30 Transition students in each aircraft and 20 Qualification students in the HH-52A will delay accumulation of stable groups in these critical areas of the training program.

DISCUSSION

It should be noted that some errors are more critical than others. Our present program does not assign differential weights to various parameters although this capability is planned for the future.

A more pressing requirement is the development of additional checkrides. We must eventually maintain four to six different automated checkrides for each aircraft in order to provide variety for returning students. It would then be possible to collect performance data on initial as well as final flights for each pair in the Proficiency Course without repeating a flight sequence.

Full implementation of this data gathering capability must wait for each new checkride to be standardized. All checkrides must then be scaled against each other on each maneuver scored. Obviously several years of development and data gathering are required before the full impact of this system will be felt. However analysis based on this present comparison tends to be encouraging.

BIBLIOGRAPHY

Caro, Paul, W., Jr. Flight evaluation procedures and quality control of training. Ft. Rucker, AL, Human Resources Research Laboratory, Division No. 6 (Aviation), Technical Report 68-3, 1968.

Caro, P. W., Jr., Hall, E. R. and Brown, CDR G. E. Design and procurement bases for Coast Guard aircraft simulators. Ft. Rucker, AL, Human Resources Research Organization, Division No. 6 (Aviation), Technical Report 69-103, 1969.

Hall, E. R., Caro, P. W., Jr., Jolley, O. B. and Brown, CDR G. E., Jr. A study of U.S. Coast Guard aviator training requirements. Ft. Rucker, AL, Human Resources Research Organization, Division No. 6 (Aviation), Technical Report 69-102, 1969.

Jensen, R. S., VanderKolk, R. J., FitzHenry, P., Rose, D. J. and Roscoe, S. N. Pilotage error in area navigation--performance of instrument pilots using symbolic displays. Washington, D.C., Federal Aviation Administration, Final Report FAA-RD-71-31, 1971.

Knoop, Patricia A. Advanced instructional provisions and automated performance measurement. *Human Factors*, 1973, 15(6), 583-597.

Koonce, Jefferson M. Effects of ground based aircraft simulator motion conditions upon prediction of pilot proficiency. Savoy, Ill., Univ. of Illinois, Urbana Champaign Institute of Aviation, Aviation Research Laboratory, Technical Report ARL-1974-5/AFOSR 74-3, April 1974.

Povenmire, H. K., Alvares, Kenneth M. and Damos, Diane L. Observer-observer flight check reliability. Savoy, Ill., University of Illinois Institute of Aviation, Aviation Research Laboratory, Technical Report LF 70-2, 1970.

Smith, J. F., Flexman, R. E. and Houston, R. C. Development of an objective method of recording flight performance. Lackland Air Force Base, Texas, Human Resources Research Center, Technical Report HRRC 52-15, December 1952.

```
'INSTRUMENT CLIMB'

  SEGMENT 1

    BLOCK 1

      SCORING CRITERIA  AIRSPEED    50,  60
                        HEADING    125, 135
                        TORQUE      90, 100

      END CRITERIA ALTITUDE GREATER THAN 2900 OR 240
      SECONDS AFTER START OF BLOCK 1.

    BLOCK 2

      SCORING CRITERIA  AIRSPEED    55,  80
                        ALTITUDE  2900, 3100
                        HEADING    125, 135

      END CRITERIA AIRSPEED GREATER THAN 70 OR 120
      SECONDS AFTER START OF BLOCK 2.

    BLOCK 3

      SCORING CRITERIA  AIRSPEED    70,  80
                        .
                        .
                        .
```

Figure 3. Format of SANSKRIT checkride program.

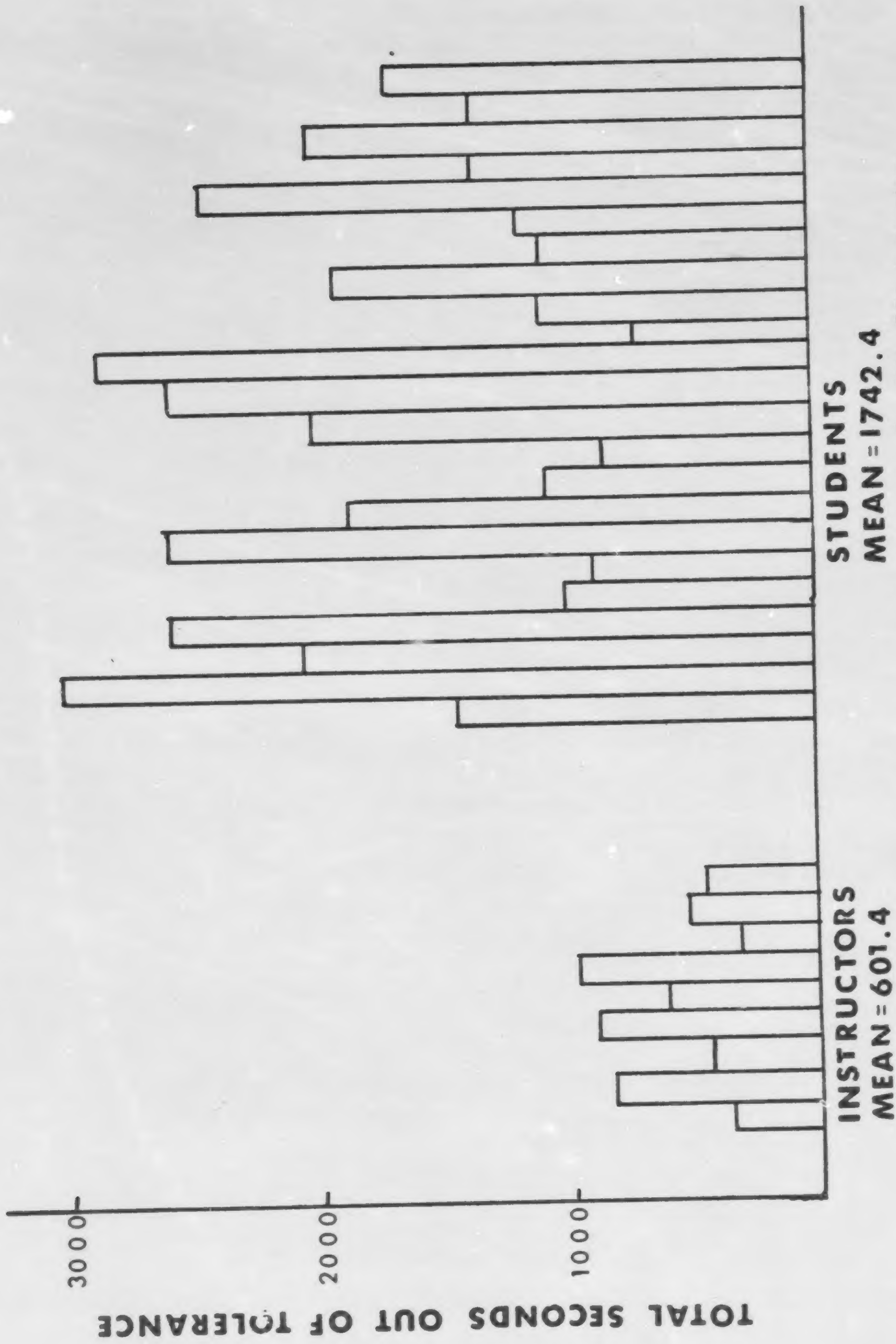


Figure 4. Distribution of automated checkride scores for HH-52A instructors and HH-52A proficiency students, ($p < .001$)

	INSTRUMENT TAKEOFF	AIRWAYS TRACKING	HOLDING PATTERN	VOR APPROACH	RADIO BCN TRACKING	ADF APPROACH	ILS APPROACH	TOTAL
INSTRUCTORS	35.0	108.1	68.3	63.7	3.7	45.9	267.8	601.4
STUDENTS	267.1	534.6	121.1	143.9	58.6	153.3	455.2	1742.9

Table 1. Mean composite maneuver scores for each maneuver. All differences significant ($p<.01$).

AN AUTOMATED AIRCREW STATION ANTHROPOMETRIC UNIT

By

Carole A. Bohn

Military aircrew stations, as currently designed have failed to consider the constraints imposed by the physical characteristics of the personnel for whom they are being constructed. The present designs are based on the "average man" (i.e. 50th percentile on all anthropometric dimensions) or some specific range of anthropometric sizes, for example, 5th through 95 percentile. Such an approach has met with several very serious restrictions. The range selection criteria has as a basic assumption that all persons are perfectly proportioned (e.g. a 30 percentile man has a 30th percentile reach, a 30th percentile sitting eye height, etc.). Moroney and Smith have shown that such an assumption is false, leading to severe reduction in the potential population of users of military aircraft. Over 50% of the potential user population tested was found to exceed the established anthropometric percentile design limits when the 5th thru 95th percentile limits were used.

The data base for designs has historically been gathered via the standard anthropometric chair. Such measurements typically do not include tolerances for various clothing or equipment configurations. Such equipment significantly alters the crewman's position with the nature of the displacement frequently dependent on anthropometric size. The base data have little relationship to the actual man-machine interface since it is gathered in an artifical and static environment.

The effects of such design failures range from simple discomfort to inability to reach critical controls requiring immediate activation. Gregoire and Barnes found anthropometric incompatibilities to be a significant problem in certain aircraft. These incompatibilities were based in part on improper design criteria. The lack of a firm data base both with respect to actual measurement criteria in the man-machine interface and the consequences of designed-in anthropometric constraints on operator performance can account for this lack of proper consideration.

The following five slides illustrate excessive reach requirements in the some of the most recently designed military tactical aircraft.

METHOD

A computer controlled and monitored anthropometric unit was designed to facilitate the gathering of this required data. A DEC PDP-8E IDACS computer equipped with a seven channel analog-to-digital converter was interfaced with a variable-geometric cockpit mockup to measure anthropometric dimensions pertinent to aircrew station design. Four measurement points were implemented. Three measurement points were installed in the instrument panel facing the subject, one in the upper right hand side of the panel, one in the upper left hand side of the panel and a third in the lower left hand side of the panel. A fourth measurement point was mounted on a sliding scale parallel to the back of the subject's chair. Buttock to knee dimensions could also be measured using lower panel indices. Each point was independently connected to a variable resistor which regulated a constant power supply. The output of this configuration was a zero to plus one volt for interface with a single analog-to-digital channel. The regulated power supply was also attached directly to a single analog-to-digital channel for use as a reference voltage. This allowed the variation in the measurement points to be converted from a length dimension to a voltage dimension for computer interface.

The analog-to-digital channels then converted the voltage to a decimal number in binary code (in the range zero to five hundred). This resulted in a per inch conversion constant of 44.11, allowing measurement accuracies to almost one fifthieth of an inch.

A computer program written in PAL8 assembly language analized the analog-to-digital channel inputs. A standard sampling rate of two-hundred-fifty per second was chosen for all channels. Data was gathered by repeated averaged samples (N=5) of a channel's input with the maximum average being taken as the final measurement input. In order to compensate for the fluctuations in the regulated power supply the measurement was converted. The power supply channel was sampled as soon as the macimum average on the measurement channel was detected. A reference average voltage was computed based on a sample of five readings. The measurement value was subtracted from this reference voltage. The resultant value was then divided by the reference voltage, yielding the final measurement as a percentage of current power supply votage. To convert this percentage figure to inches it was multiplied by a measurement specific constant. The measurement specific constant was formed by dividing the maximum channel range by the per inch conversion factor as given above. The channels were sampled consecutively, conversions being computed on a real-time basis.

The system was designed with the ability to be modified from the standard measurement configuration to a realistic crew station configuration. The system operation, after the configuration had been selected, required only that the subject be seated properly and sequentially depress the measurement points directly in front of him. The single vertical measurement point was then raised by the experimenter to the desired height to be measured. The per subject measurement time was under two minutes for the following thirteen anthropometric dimensions:

1. Erect sitting eye height
2. Erect sitting eye depth
3. Erect functional reach
4. Functional sitting eye height
5. Functional sitting eye depth
6. Functional non-stretching reach
7. Functional maximum stretching reach
8. Static buttock to knee length
9. Function buttock to knee length
10. Difference between erect sitting eye height and functional sitting eye height
11. Difference between erect sitting eye depth and functional sitting eye depth
12. Difference between functional maximum stretching reach and functional non-stretching reach
13. Difference between erect functional reach and functional non-stretching reach

DISCUSSION

The computerization of the newly designed facility provides an environment where needed data can be collected on critical anthropometric problems. The variable-configuration nature of the system allows for collection of data on the relationship between the standard measurement technique and measurements taken in a more realistic environment. The displacement of personal and survival equipment could thus be quantified and used in future design standards.

The automation of the measurement procedure not only provides more accurate data with minimial operator training, but creates the potential for an anthropometric data bank to be used in the development of design criteria for crew stations.

The variability of the system establishes a unit which can be used as a personnel selection tool based on current designed-in anthropometric constraints and a device to determine basic anthropometric constraints which must be considered in the design of the crew station configuration.

The most significant advance provided by the new system is the ability to delimit the relationship between performance and anthropometric variables. Currently the unit is equipped to present auditory vigilance and/or tracking tasks, and to take reaction time measurements to visual and/or auditory stimuli.

The computerization of the newly designed facility provides a significant state of the art advancement in the capability of evaluating aircrew station modifications before and during design rather than after the fact. Additionally, this facility can provide an automated and accurate establishment of anthropometric design criteria based on realistic performance data.

MLM RESEARCH REPORT 74-1

PREDICTION FROM CONTINGENCY TABLES USING JOINT LIKELIHOODS

Lieutenant-Commander W.S. Shields

The research for this paper was supported in part by the Defence Research Board of Canada, Grant Number 9465-12. The views expressed are those of the author, and do not necessarily reflect official views or policy of the Royal Military College or the Department of National Defence.

Department of Military Leadership and Management
Royal Military College of Canada, Kingston, Ontario
Department Head: Lieutenant-Colonel G.J. Carpenter

PREDICTION FROM CONTINGENCY TABLES USING JOINT LIKELIHOODS

Lieutenant-Commander W.S. Shields
Department of Military Leadership and Management
Royal Military College of Canada

A procedure for predicting categorical outcomes using categorical predictor variables was described by Moonan (1972). This paper describes a related technique which uses prior probabilities, updated by joint likelihoods, as classification criteria. The procedure differs from Moonan's in that the outcome having the greatest posterior probability is selected as the prediction regardless of misclassification cost. It also differs in method of screening and weighting the predictor variables, and treats the problem of small-sample bias. Applications, to date, are in the analysis and use of questionnaire responses to predict categorical outcomes, —namely voluntary, academic, and military attrition from a Service College. Classification efficiency appears to be comparable to that of the Moonan technique.

Problem

The most common source of categorical data in behavioral research is the questionnaire. Whether questions are multiple-choice, or responses are grouped after-the-fact, it is usually difficult and frequently impossible to order responses along a metric scale. Even when an array of choices has a metric design, nonlinearity of relationships may make it preferable to treat responses as qualitative. The usual purpose of the questionnaire is to try to predict some criterion variable. Like the predictor variables (questionnaire responses) the criterion variable may be either quantitative (metric) or qualitative (categorical).

If either the criterion variable or its predictors are metric, discriminant analysis or a related technique may be used, unless relationships are nonlinear. If neither is metric some form of categorical analysis must be used. One option is to treat each category of each variable as a separate zero-one variable. The difficulty with this is proliferation in the number of variables. If each question has 5 possible responses and one asks 100 questions, 500 variables result. These have so many possible interactions that ones resulting from errors of measurement commonly dominate the analysis.

Moonan (1972) pioneered a strategy of analysis which treats the responses of a candidate to a selected subset of questions as a single unit and calculates the Bayesian conditional probability of an outcome to be predicted, given the candidate's particular response profile, under an assumption of independence of the questionnaire items. A classification by outcome is then made which minimizes total misclassification cost.

Method

The strategy reported here is related to the first part of Moonan's procedure. It computes the joint likelihood of the candidate's responses under an assumption that a given outcome will occur, and multiplies this by the prior probability of the outcome, thus obtaining a quantity proportional to its posterior probability[1]. The outcome having the greatest posterior probability is then selected as the prediction. Because the joint likelihood of a given outcome, under an assumption of independence, is the product of the individual likelihoods derived from the predictor questions for a given set of responses, one may use the sum of the log-likelihoods as a sufficient statistic. The logarithm of the prior probability is then added to this sum and the result compared with that of other outcomes.

Because the concept of "likelihood" is used more often with continuous than with categorical variables, it should be defined carefully here. Suppose that an entire population could be entered into a contingency table such as that in Figure 1.

Figure 1

	a_1	a_2	a_3	
B_1	f_{11}	f_{12}	f_{13}	b_1
B_2	f_{21}	f_{22}	f_{23}	b_2
B_3	f_{31}	f_{32}	f_{33}	b_3
B_4	f_{41}	f_{42}	f_{43}	b_4
	A_1	A_2	A_3	

---

[1] This procedure was prescribed by Jeffreys (1939). He wrote: (p. 29) "The posterior probabilities of the hypotheses are proportional to the products of the prior probabilities and the likelihoods." Later (p. 133) he added "... when several estimates are combined the part from the prior probability enters only once, while that from the likelihood enters every time."

The B_i are responses to a question and the A_j are outcomes to be predicted. The frequency in each cell is f_{ij}; the row totals are b_i, the column totals a_j. The likelihood of outcome A_j given that a candidate has made response B_i is f_{ij}/a_j. This definition is in accord with that of Jeffreys, and differs from that of R.A. Fisher only in that the likelihood is required to *equal* the probability of the event which has been observed, given that the hypothesis under consideration is true, and not merely be proportional to it. This distinction makes it possible to perform any operation on the elements of a column vector, when considering likelihoods, that one would perform on a row vector when considering probabilities.

Independence

Before proceeding, some justification should be given for the assumption of independence of the categorical predictor variables. Moonan (1973) justifies it partly on the basis of computational feasibility. Certainly much less computer storage is required under this assumption, because only contingency tables relating to the outcome to be predicted need be stored, rather than a set which includes every predictor versus every other predictor. More important than this, if joint likelihoods were to be inferred directly from the data, the criterion sample would have to be extremely large, unless the number of questions were extremely small, for a sufficient cell content in the multi-dimensional contingency table that one could trust the results. For example, if 20 questions were to be used as predictors, a cell content greater than unity would exist only if two persons in the criterion sample had answered all 20 questions in exactly the same way, — a rare event indeed.

Another argument for the assumption of independence is that the consequences of failure of this condition are rarely serious. First of all, it is unusual for categorical variables to be intercorrelated as highly as metric ones. Commonly, one or two categories of a variable will correlate highly with one or two categories of another variable, ("Province of Birth in Canada" versus "Mother Tongue", for example) but overall correlations are usually low, unless one has accidentally (or deliberately) asked the same question twice.[2] Secondly, the effect of such correlation is merely to give some additional weight to highly correlated questions. Questions asked twice, for example, receive double weight. If this prejudices the prediction, it may do so in a favourable way because if a researcher has asked many questions centered in a particular area, or the same question in a number of different ways, it is presumably because he believes this area or question to be important.

---

[2] Moonan (1972) also used these arguments. He wrote "... many qualitative characters in practical problems are likely to be nearly independent and their dependencies poorly estimated."

Selection of Predictors

χ^2 is a convenient criterion of variable selection because it is easily tested for significance. The author prefers the likelihood measure

$$\chi^2 \simeq -2 \ln \lambda$$

over the Pearson χ^2 measure because of its stability when as many as 50% of the contingency table cells — not counting entirely blank rows and columns — are vacant. Wilks (1928) demonstrated that $-2 \ln \lambda$ is just as trustworthy as the Pearson approximation. The measure is identical to that derived from Shannon information theory:

$$\chi^2 \simeq 2N\hat{T}$$

$$\text{where } \hat{T} = \hat{H}(A) + \hat{H}(B) - \hat{H}(A, B)$$
$$= \hat{H}(A) - \hat{H}(A|B)$$
$$= \hat{H}(B) - \hat{H}(B|A)$$

and N is the sample size. A and B are two categorical variables, and the uncertainties (entropies) $\hat{H}$ are calculated in "nits" using:

$$\hat{H}(A) = -\sum_{j} \frac{a_j}{N} \ln \frac{a_j}{N}, \quad \hat{H}(B) = -\sum_{i} \frac{b_i}{N} \ln \frac{b_i}{N}, \text{ and}$$

$$\hat{H}(A, B) = -\sum_{ij} \frac{f_{ij}}{N} \ln \frac{f_{ij}}{N}$$

Having found a subset of predictors possessing significant relationships with the outcome to be predicted, it is suggested that the contribution of each predictor to the total log-likelihood be given a weight proportional to Newman and Gerstman's (1952) coefficient of constraint:

$$\hat{D}(A|B) = \frac{\hat{T}}{\hat{H}(A)} = \frac{\hat{H}(A) - \hat{H}(A|B)}{\hat{H}(A)}$$

Giving inferior weight to questions of low relevance can be justified on the basis of enhancing the "signal-to-noise" ratio. $\hat{D}(A|B)$ is the relative reduction in the uncertainty of A when B is known, and is asymmetrical. If desired, $\hat{D}$ can be corrected for bias using formulas developed by Miller (1954) for correction of $\hat{H}$ and $\hat{T}$:

$$H \simeq \hat{H} + (K - 1)/2N$$

where K is the number of occupied categories in the variable, and

$$T \simeq \hat{T} - (R - 1)\ (C - 1)/2N$$

where R is the number of occupied rows, and C the number of occupied columns in the contingency table. (K-1) and (R-1) (C-1) will be recognized as the number of degrees of freedom appropriate to the x^2 test. Negative values of T are set to zero.

When predictors are weighted proportional to D or $\hat{D}$, it is the practice of the writer to weight the prior distribution equally with the strongest predictor. This amounts to giving each a weight of unity, because a constant factor will not affect the comparison of a number of posterior probabilities. The decision to weight the prior in this manner is supported by the fact that it is based on a sample of the same size as is each predictor. The prior is simply $\ln\ (a_j/N)$.

Bias

Suppose k cell members occur out of a possible (column total) n. The "maximum-likelihood" estimate of the likelihood is k/n. This estimate, although bias-free for addition, has negative bias for multiplication (addition of log-likelihoods). It was decided to seek a formula for multiplicative use which would be as free as possible of bias over a broad range of population likelihoods.

If the sampling method resembles a Bernoulli process, and p represents the true population likelihood, two consecutive cell frequencies r and r+1 will have the same expectation when

$$\frac{n!}{r!(n-r)!}\, p^{r}(1-p)^{n-r} = \frac{n!}{(r+1)!\ (n-r-1)!}\, p^{r+1}\ (1-p)^{n-r-1}$$

or when $p = \frac{r+1}{n+1}$. These will also be the two most commonly observed cell frequencies, and will contain bias of opposite sign. If these biases are to be mutually cancelling, then the likelihood estimates $\hat{p}$ should be such that

$$\hat{p}_r \hat{p}_{r+1} = p^2$$

Letting $\hat{p}$ have the form $\frac{k+a}{n+b}$, the above becomes:

$$\left(\frac{r+a}{n+b}\right)\left(\frac{r+1+a}{n+b}\right) = \left(\frac{r+1}{n+1}\right)^2$$

Because r and n will vary somewhat independently, we may set b=1 and solve for a, obtaining:

$$a = \sqrt{r^2+2r+1.25} \quad -r-.5$$

Unfortunately, r is unknown. However a is very nearly .5 over a broad range of values of r and approaches .5 as r increases. A test of the formula $p = (k+.5)/(n+1)$ reveals that its performance for large values of k is improved if it is modified slightly to:

$$p = \frac{k+.5}{n+.5} \quad \text{-----------------------------------} 1$$

This has very little effect on its performance for small values of k. Table 1 compares the expected value of joint likelihoods, calculated from random samples using formula 1, with population values. The estimated likelihoods are seriously biased only for small values of both p and n. In these instances sample likelihoods are bound to be poorly estimated. The bias is in the direction of avoiding the rejection of a hypothesis purely on the basis of a very small sample.

Formula 1 was adopted for estimating *both* likelihoods and prior probabilities, and was found to perform slightly better than the traditional k/n.

Missing Data

The method of this paper lends itself to a convenient and profitable treatment of missing data. One category of each variable is reserved for missing responses (or outcomes, as appropriate). The missing data category sometimes conveys important predictive information. In a study relating responses on a dental questionnaire to clinical dental examinations, the response most indicative of an unsatisfactory oral environment was refusal (or inability) to answer some of the questions.

Confidence

Because a posterior log-probability is calculated for each outcome, the difference in this quantity for the two most probable outcomes, less the difference in their priors, equals the logarithm of their likelihood ratio. This likelihood ratio provides an excellent indication of the confidence with which a given prediction can be made.

Test of the Method

The above procedure was developed for predicting categorical outcomes of Cadets at a Military College, such as academic failure, voluntary resignation, military failure, achievement of distinction, etc., based on a questionnaire written on their first day at the College. A total of 596 Cadets from four College years are currently under study. Their graduation years range from 1974 to 1978.

Rather than report results prematurely, the method will be demonstrated instead using R.A. Fisher's (1950, p 32.180) Iris data. It is chosen partly because it has already been used by several researchers, including Moonan (1972), and is widely available to future users for purposes of comparison.

Categories identical to those defined by Moonan were used in grouping the data:

Category:	1		2		3
Sepal Length:	SL	$< 4.9 \leq$	SL	$\leq 5.7 <$	SL
Sepal Width:	SW	$< 2.3 \leq$	SW	$\leq 3.0 <$	SW
Petal Length:	PL	$< 3.3 \leq$	PL	$\leq 4.9 <$	PL
Petal Width:	PW	$< 1.0 \leq$	PW	$\leq 1.5 <$	PW

Moonan used all 150 Iris plants as both criterion and prediction samples. As he pointed out, this stacks the odds heavily in favour of the classification algorithm. His program produced 9 misclassifications of the 150 Iris plants compared with 7 misclassifications made by using the same procedure with the method of this paper. The difference is nonsignificant; however, the result supports an opinion that the algorithms are of comparable quality. The method of this paper was tested also by shuffling the data cards and using the first 75 as a criterion sample from which to calculate prior probabilities and joint likelihoods for the classification of the remaining 75 plants. This resulted in 69 correct classifications and 6 incorrect ones. Misclassifications occurred only between Iris versicolor and Iris virginica.

Summary

A procedure has been described which, like Moonan's, represents a radical departure from conventional questionnaire analysis. The formula developed for making unbiased estimates of joint likelihood is equally applicable to the calculation of sample-derived joint probabilities. There are several areas in which the strategy is open to further refinement, particularly in regard to the "independence" assumption.

The comparison of hypotheses through the calculation of joint likelihoods is somewhat analogous to putting them through a long filter. A single very low likelihood, anywhere along the length of the filter, can cause a hypothesis to become "clogged" and fall hopelessly behind its competitors. Consider, for example, the following flow of information and its effect on the likelihoods of three competing hypotheses:

Hypothesis:	Mammal	Bird	Fish
Information Flow	Approximate Likelihood		
a. It has a head.	1.0	1.0	1.0
b. It has eyes.	1.0	1.0	1.0
c. It can fly.	.01	.99	.001
d. It has no feathers.	.01	0	.001
e. It has no fur.	0	0	.001

One can now classify the subject as a "flying fish" with reasonable confidence. That the choice is "unlikely" is not nearly as important as the fact that it is many times more likely than any of the available alternatives. The procedure is well summarized by a statement made by "Inspector Maigret" on a radio mystery program by that name some two decades ago. Asked the secret of his uncanny success, the famous detective replied: "Having eliminated all of the possibles, whatever remains - however improbable - must be the truth".

Table 1

EXPECTED VALUE OF A SAMPLE-DERIVED JOINT LIKELIHOOD ESTIMATE
USING THE FORMULA (K+.5)/(N+.5) FOR ITS FACTORS

SAMPLE	POPULATION LIKELIHOOD								
SIZE	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
5	0.14906	0.22228	0.30733	0.40041	0.49811	0.59797	0.69853	0.79914	0.89963
10	0.11452	0.20221	0.29914	0.39891	0.49921	0.59948	0.69967	0.79980	0.89991
15	0.10515	0.19964	0.29928	0.39951	0.49968	0.59978	0.69986	0.79991	0.89996
20	0.10180	0.19945	0.29958	0.39974	0.49983	0.59988	0.69992	0.79995	0.89998
25	0.10049	0.19957	0.29975	0.39984	0.49989	0.59993	0.69995	0.79997	0.89999
30	0.09998	0.19970	0.29983	0.39989	0.49993	0.59995	0.69997	0.79998	0.89999
35	0.09980	0.19978	0.29988	0.39992	0.49995	0.59996	0.69997	0.79998	0.89999
40	0.09975	0.19984	0.29991	0.39994	0.49996	0.59997	0.69998	0.79999	0.89999
45	0.09976	0.19988	0.29993	0.39995	0.49997	0.59998	0.69998	0.79999	0.90000
50	0.09978	0.19990	0.29994	0.39996	0.49997	0.59998	0.69999	0.79999	0.90000
55	0.09981	0.19992	0.29995	0.39997	0.49998	0.59998	0.69999	0.79999	0.90000
60	0.09984	0.19994	0.29996	0.39997	0.49998	0.59999	0.69999	0.79999	0.90000
65	0.09986	0.19995	0.29997	0.39998	0.49998	0.59999	0.69999	0.80000	0.90000
70	0.09988	0.19995	0.29997	0.39998	0.49999	0.59999	0.69999	0.80000	0.90000
75	0.09990	0.19996	0.29998	0.39998	0.49999	0.59999	0.69999	0.80000	0.90000
80	0.09991	0.19997	0.29998	0.39999	0.49999	0.59999	0.70000	0.80000	0.90000
85	0.09993	0.19997	0.29998	0.39999	0.49999	0.59999	0.70000	0.80000	0.90000
90	0.09993	0.19997	0.29998	0.39999	0.49999	0.59999	0.70000	0.80000	0.90000
95	0.09994	0.19998	0.29999	0.39999	0.49999	0.60000	0.70000	0.80000	0.90000
100	0.09995	0.19998	0.29999	0.39999	0.49999	0.60000	0.70000	0.80000	0.90000
105	0.09995	0.19998	0.29999	0.39999	0.49999	0.60000	0.70000	0.80000	0.90000
110	0.09996	0.19998	0.29999	0.39999	0.49999	0.60000	0.70000	0.80000	0.90000
115	0.09996	0.19998	0.29999	0.39999	0.50000	0.60000	0.70000	0.80000	0.90000
120	0.09997	0.19999	0.29999	0.39999	0.50000	0.60000	0.70000	0.80000	0.90000
125	0.09997	0.19999	0.29999	0.39999	0.50000	0.60000	0.70000	0.80000	0.90000
130	0.09997	0.19999	0.29999	0.39999	0.50000	0.60000	0.70000	0.80000	0.90000
135	0.09997	0.19999	0.29999	0.40000	0.50000	0.60000	0.70000	0.80000	0.90000
140	0.09998	0.19999	0.29999	0.40000	0.50000	0.60000	0.70000	0.80000	0.90000
145	0.09998	0.19999	0.29999	0.40000	0.50000	0.60000	0.70000	0.80000	0.90000
150	0.09998	0.19999	0.29999	0.40000	0.50000	0.60000	0.70000	0.80000	0.90000

REFERENCES

Bayes, Thomas. "An Essay Towards Solving a Problem in the Doctrine of Chances" (1773). In: Deming, W.E. *Facsimiles of Two Papers by Bayes.* New York: Hafner Publishing Co., 1963.

Bowser, S.E. "Applications of Predictor Ordering and Selection by a Bayesian-Decision Technique", *Proceedings of the 1973 Military Testing Association Conference,* San Antonio, Texas, 28 October - 2 November, 1973.

Edwards, A.W.F. *Likelihood.* Cambridge: University Press, 1972.

Fienberg, S.E. and Holland, P.W. "Simultaneous Estimation of Multi-nominal Cell Probabilities", *Journal of the American Statistical Association,* 68 No. 343 (September, 1973), 683-91.

Fisher, R.A. *Contributions to Mathematical Statistics.* New York: John Wiley and Sons, Inc., 1950.

__________. *Statistical Methods and Scientific Inference.* Edinburgh: Oliver and Boyd, 1956.

Good, I.J. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* Cambridge, Mass: The M.I.T. Press (Research Monograph No. 30), 1965.

Harville, D.A. "Assigning Probabilities to the Outcomes of Multi-Entry Competitions", *Journal of the American Statistical Association,* 68 No. 342 (June, 1973), 312-16.

Jeffreys, H. *Theory of Probability.* Oxford: Clarendon Press, 1939.

Laplace, Marquis de. *Theorie anatytique des probabilites.* Paris: Courcier, 1820.

Miller, George A. "Note on the Bias of Information Estimates", in Quastler, H. (Ed.), *Information Theory in Psychology,* Glencoe, Ill.: The Free Press, 1955.

Moonan, W.J. "ABCD: A Bayesian Technique for Making Discriminations with Qualitative Variables", *Proceedings of the 1972 Military Testing Association Conference,* Lake Geneva, Wisconsin, 18-22 September, 1972.

__________. "Charosel: A Computer Program which Selects Qualitative Predictors for Qualitative Criterion Prediction Problems", *Proceedings of the 1973 Military Testing Association Conference,* San Antonio, Texas, 28 October - 2 November, 1973.

Newman, E.B. and Gerstman, L.J. "A New Method for Analyzing Printed English", *Journal of Experimental Psychology*, 44 (1952), 114-25.

Shields, W.S. "The Use of Qualitative Information in the Prediction of College Attrition", *Proceedings of the First Annual Conference, Canadian Association of Administrative Sciences*, Queen's University, Kingston, Ont., 31 May - 1 June, 1973.

__________. "The Use of Information Tehory in the Selection and Weighting of Cluster Variables for Prediction", *Proceedings of the Southeastern Regional Meeting, The Institute of Management Sciences*, Atlanta, Georgia, 18-19 October, 1973.

Wilks, S.S. "The Likelihood Test of Independence in Contingency Tables", *Biometrika*, 20A (1928), 263-94.

SECTION VII

TRAINING METHODOLOGIES

283

G-BBE-2
23 September 1974

BOATING SAFETY EDUCATION: THREE ROUTES TO A COMMON GOAL

The U. S. Coast Guard, Office of Boating Safety has a training problem. Although we have devised a solution, the solution itself has presented us with some additional problems. The purpose of this paper is to explain the problem and the solution in hopes that we can learn how to solve the problems our solution has created.

The initial problem is this: How do we meet the training needs of three different types of students? To solve this problem, the Coast Guard has devised three different training programs. As a result of our solution, however, we have a new question to answer: Do the students in all three courses learn the same thing? The investigation of our training and testing was to answer this question.

First of all, we defined the three types of students that we have to train. First, we have the student whose full time job is boating safety. Some are Coast Guardsmen - regular and Reserve - others are federal, state and local boating law officers and administrators. This student needs comprehensive, intensive instruction, and, because boating safety is his profession, he can devote himself to full time training. To meet his needs and to make the most of his availability for training, the Coast Guard established a resident school, the National Boating Safety School (NBSS). This school is one part of our three-sided solution.

The second type of student we have works full or part time at boating safety. The personnel are much the same as those who attend NBSS - regular and Reserve Coast Guard, federal, state and local officials. The major difference is in the limited amount of time he needs or is able to devote to the training. For this student, the Office of Boating Safety has established a 32-hour course, boating safety scaled down to the bare necessities. This course is taught in each of the 12 Coast Guard Districts, usually by graduates of NBSS. The student needs fast, efficient qualification for duty as a boating safety boarding officer; thus, the 32-hour course provides the second part of the solution for boating safety training.

The third type of student cannot be spared from his regular work, whether it is boating safety or not. This student must rely on a correspondence course to learn about boating safety. For him, the ideal solution is the Boating Safety or BOSAF Course, which is prepared at the U. S. Coast Guard Institute.

Each of the solutions - NBSS, the 32-hour course, the BOSAF Course - suits student needs for subject matter and study requirements. Upon successful completion of any one of the courses, and with the recommendation of the student's Commanding Officer, the student receives a D-8 qualification code. This qualification identifies a person as a boating safety boarding officer, legally entitled to board a recreational boat and perform certain tasks as defined in the U. S. Coast Guard Enlisted Ratings Qualification Manual (CG-311). At the end of each of the three courses of study, a student takes some sort of final examination. Our question, re-stated in more specific terms, is this: Are all the students from each type of training being tested on the same facts and procedures which are considered necessary for the duty of a boarding officer? In other words, are all students tested on the same kernel of knowledge?

To determine the answer, we began by setting down the areas of knowledge we felt a person should have if he is to be validly qualified as a boarding officer. We found five general areas:

1. Federal and state boating law requirements.
2. Boat examination procedures.
3. Law enforcement responsibilities.
4. Boating accident reporting procedures.
5. Coast Guard Auxiliary functions.

Once we had stated these five areas, we translated them into general job tasks. This helped us compare the action of the job task with the action being tested. Translated into general job tasks, the general areas of knowledge are these:

1. Federal and state boating law requirements.
    a. Explain the requirements to the public.
    b. Apply the requirements to practical situations.

2. Perform boat examinations.
3. Enforce federal boating laws.
4. Explain the process for reporting boating accidents to the public.
5. Describe the U. S. Coast Guard Auxiliary to the public.

    a. History
    b. Organization
    c. Functions.

Our next step was to review the final exams from each course and label each test item according to the area of knowledge it most logically fit. Because the 32-hour course is taught in each district, the instructors sometimes adapt items to fit regional situations. Therefore, we reviewed both the standardized 32-hour exam and a specialized 32-hour exam from the Second District. On all the exams some items included facts from more than one area of knowledge and were labeled accordingly. For example, one test item could deal with both federal requirements and boat examination procedures. Other items did not fit under any of the necessary general areas; these items were labeled "no" to indicate that they were not required for a D-8 boarding officer to perform adequately. Any items in this category, therefore, fell outside the common kernel of knowledge we were looking for.

Our results showed that the kernel knowledge items made a percentage large enough that, if the student knew all the items, he could pass the exam. In other words, a passing percentage of each exam was devoted to items required for the D-8 qualification.

Exam	% Kernel Items	% Passing
NBSS	81%	70%
Standardized 32-hour	91%	70%-80%
2nd District 32-hour	92%	80%
BOSAF	84%	80%

Although the kernel of knowledge does appear to be tested adequately in each case, several questions concerning the validity of the training program and the validity of our review process should be asked. Possibly, we can find some of the answers from members of the Military Testing Association.

1. The three courses have different passing requirements:

NBSS	70%
32-hour	70%-80% (Varies by district)
BOSAF	80%

Does this make any difference in the number of items we should require as kernel testing items?

2. Was our review procedure too subjective? Whether or not items should have been included as kernel items could have been, in some cases, a matter of opinion. We know we discovered some useful facts concerning our three courses, but we would like to know if we adequately answered the questions we set out to answer. If we did not, we need to set about the review another way so that the qualifications of a D-8 boarding officer will be as consistent as possible.

3. To ensure consistency, should the Office of Boating Safety consider building a block of kernel knowledge questions that could be adapted to the tests given at the end of each course? Would this be too restrictive to the various training units involved?

4. One fact that our review showed was that the distribution of items among the general knowledge areas was lopsided. Kernel Area One, federal and state requirements, applied to many test items, either alone or in combination with another kernel area. For this reason, Kernel Area One dominated the test items and other areas were shortchanged. Therefore, we felt some redistribution of items might be useful in testing the kernel knowledge more adequately. The distribution of items was as follows:

NBSS Final

Kernel Area	1:	52 items
	2:	36 items
	3:	6 items
	4:	4 items
	5:	1 item

Standardized 32-Hour Course Examination

Kernel Area 1: 70 items
2: 25 items
3: 23 items
4: 1 item
5: 2 items

2nd District 32-Hour Course Examination

Kernel Area 1: 70 items
2: 20 items
3: 19 items
4: 3 items
5: 2 items

BOSAF (51)

Kernel Area 1: 64 items
2: 20 items
3: 16 items
4: 6 items
5: 10 items

BOSAF (52)

Kernel Area 1: 60 items
2: 20 items
3: 13 items
4: 4 items
5: 4 items

BOSAF (53)

Kernel Area 1: 71 items
2: 6 items
3: 4 items
4: 1 item
5: 5 items

The question here is, how do we determine how to redistribute? Is this again purely a matter of opinion?

Our solution to the three-sided problem of gearing our boating safety training to three different types of students posed the initial problem of adequate subject coverage and consistency. The review and investigation has revealed several other problems of coverage and consistency that we ought to surmount. We still feel our three routes to a common goal of a boarding officer, is a valid training approach. Our ambition is to make the routes as accurate as possible so that the D-8 qualification, no matter how it is attained, reflects an accurate picture of the abilities and skills one should expect to find in a qualified boating safety boarding officer.

STUDENT RESPONSE IN TELEVISION INSTRUCTION

Fenton W. Brashear

ABSTRACT

The objective of this study was to investigate the effectiveness of passive, covert and overt methods of student response in individual television instruction. Students were classified on the basis of high and low achievers, then randomly assigned to one of four treatment groups. Group I (passive) was given conventional television instruction without opportunity to respond during the instruction. Group II (covert) was given multiple-choice recall questions during the instruction and asked to make a mental selection of the correct answers. Groups III and IV (both overt) were also given questions during the instruction and asked to actively select the correct answers. A two-way analysis of variance of the criterion test scores revealed significant differences on both the posttest and the delayed retention test. On the posttest, significant differences were found between the passive group and each of the three student response groups in both achievement levels. However, no significant differences were found among the three student response groups. On the retention test given one week later, low achievers in the covert group obtained significantly higher scores than the low achievers in the passive group. These findings, in favor of student response in individual television instruction, support the related research which has generally found student response superior in group instruction.

Numerous experimental studies have been conducted using various methods of student response during group instruction. Most of these studies have utilized instructor presentations, training films and television instruction. Related literature reveals very little research involving student response in individual or self-paced instruction with the exception of programed instruction. The students participating in this experimental study received individual television instruction with particular interest on the effectiveness of conventional passive, covert and overt methods of student response during instruction.

PROCEDURES

The learning conditions for the study were provided by a student carrell equipped with a television monitor and headset. A video tape was produced from a lesson in the Automotive Repairman (63H20) Course at the U. S. Army Ordnance Center and School, Aberdeen Proving Ground, Maryland. Modified versions were prepared from the basic video tape to provide for four treatment methods. Group I (passive) students received the subject matter of the lesson with intermittent summaries of the major teaching points during the presentation. Group II (covert) students received the same subject matter presentation plus the addition of twelve multiple-choice recall questions at intervals within the television presentation. Students in this group were instructed to review the choices presented as each question appeared and to "think" the correct choice. After a pause, the correct answer was given by the tape. Group III (pencil/paper) students received the same subject matter and questions but were instructed to mark the correct answer on a test sheet as each question was presented. After a pause, they were given the correct answer. Group IV (responder unit) students were given the same instruction and questions but were instructed to push a button (A, B, C or D) on a responder unit to indicate their choice. After a pause, they were also given the correct answer, or remedial instruction by the tape if they had selected the incorrect answer.

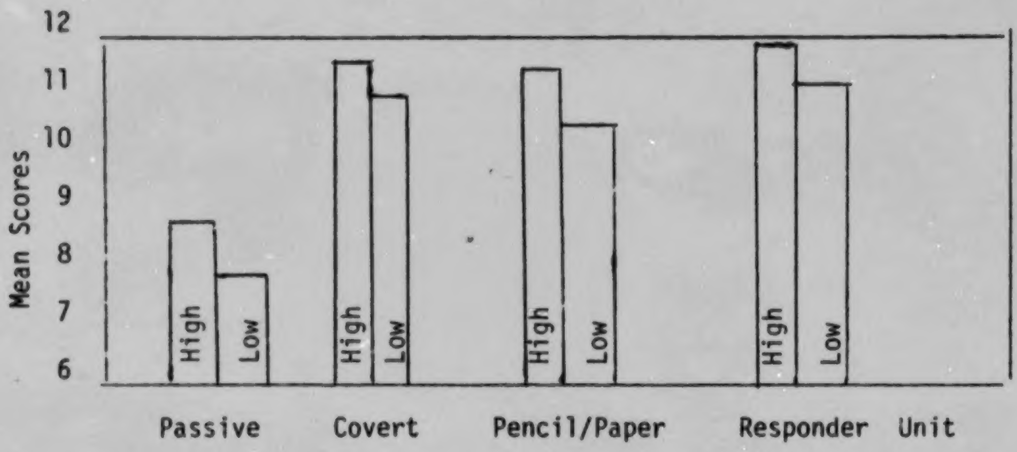
In addition to the methods variable (of four categories above) the students were classified on the basis of high and low achievers. Selection was determined by the relation of each student's cumulative achievement score in the course just prior to the experimental instruction to the class mean. High and low achievers were then randomly assigned to one of the four method groups. This provided two independent variables for study (methods of instruction and achievement levels) and resulted in a 4 x 2 mixed effects factorial experimental design based on the four methods of instruction and two levels of achievement.

DATA COLLECTION AND ANALYSIS

The criterion scores for evaluation were the number of correct responses to the same twelve questions that were presented during the experimental instruction on the posttest and the delayed retention test. Scores of 112 students in eight equal cell entries were provided for the evaluation of the posttest which was administered immediately following the experimental instruction. The retention test, given one week later, provided 80 useable scores in the eight equal cells for evaluation. Data analysis was made using a two-way analysis of variance. This analysis permitted the segregation of the two independent variables so that any significant differences would be revealed and any interaction variances identified.

RESULTS

The results of the posttest are graphically shown in Figure 1.



Posttest Methods of Instruction and Achievement Levels
Figure 1

Both high and low achievers of the passive method were much lower than the means of the three response methods. High achievers exceeded low achievers in all four methods of instruction.

The statistical analysis revealed significant differences in the post-test data as shown in the two-way analysis of variance summary Table 1 below.

Table 1. Posttest Analysis of Variance Summary Table

Mixed Design Source	d.f.	Mean Square	F-Ratio	P
Main Effect				
Methods of Instruction	3	64.3571	24.50	.0005
Achievement Levels	1	26.0357	128.65	.001
Interaction				
Methods of Instruction Achievement Levels	3	0.2024	.08	.972
Within cells	104	2.6264		

The variance among the methods of instruction was highly significant at the .0005 level of probability. The achievement levels also resulted in a very significant difference at the .001 probability level. No interaction variances were found on the posttest results since the F-ratio was less than 1.

Since significant differences were found in the independent variables, further analysis was necessary to identify the sources of variation. This was accomplished by the comparison of pairs of means to determine which pair or how many pairs contributed to the significant differences within each variable. The Scheffe value of minimum significance test was applied to identify pairs of means having significant differences. In Table 2 are six of twelve possible comparisons found to have significant differences at acceptable levels of probability.

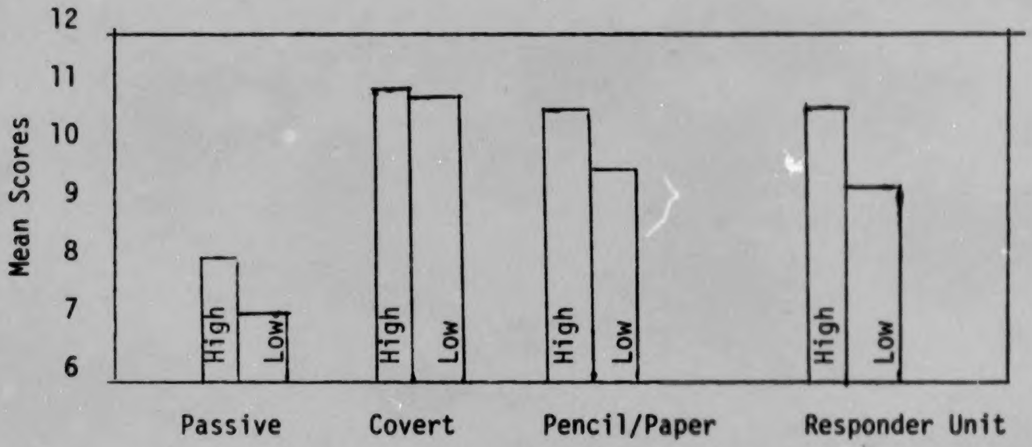
Table 2. Significant Differences of Achievement Levels and Methods of Instruction

Level	Methods Compared	Mean Difference	P
High Achievers	Passive-Covert	-2.858	.005*
	Passive-Pencil/Paper	-2.715	.010*
	Passive- RU	-3.215	.001*
Low Achievers	Passive Covert	-3.214	.001*
	Passive-Pencil/Paper	-2.714	.010*
	Passive-RU	-3.286	.001*

*Probability at an acceptable level of significance.

The significant differences above indicated that each of the student response methods (covert, pencil/paper and RU) were superior on the posttest to the passive method for both high and low achievers. However, no significant differences were found in the comparison of any pairs of the three student response methods.

The results of the retention test are graphically shown in Figure 2.



Retention Test Methods of Instruction and Achievement Levels
Figure 2

As in the posttest, the retention test means on both achievement levels of the passive method were considerably below the means of each of the three student response methods. Although high achievers exceeded low achievers in all four methods, as in the posttest, the mean difference in the covert achievement levels on the retention test above was exceedingly small (.100).

The statistical analysis revealed significant differences in the retention test data as shown in the two-way analysis of variance summary Table 3 below.

Table 3. Retention Test Analysis of Variance Summary Table

Mixed Design Source	d.f.	Mean Square	F-ratio	P
Main Effect				
Methods of Instruction	3	44.9125	14.444	.0005
Achievement Levels	1	19.0125	10.69	.047
Interaction				
Methods of Instruction Achievement Levels	3	1.7792	.57	.635
Within cells	72	3.1097		

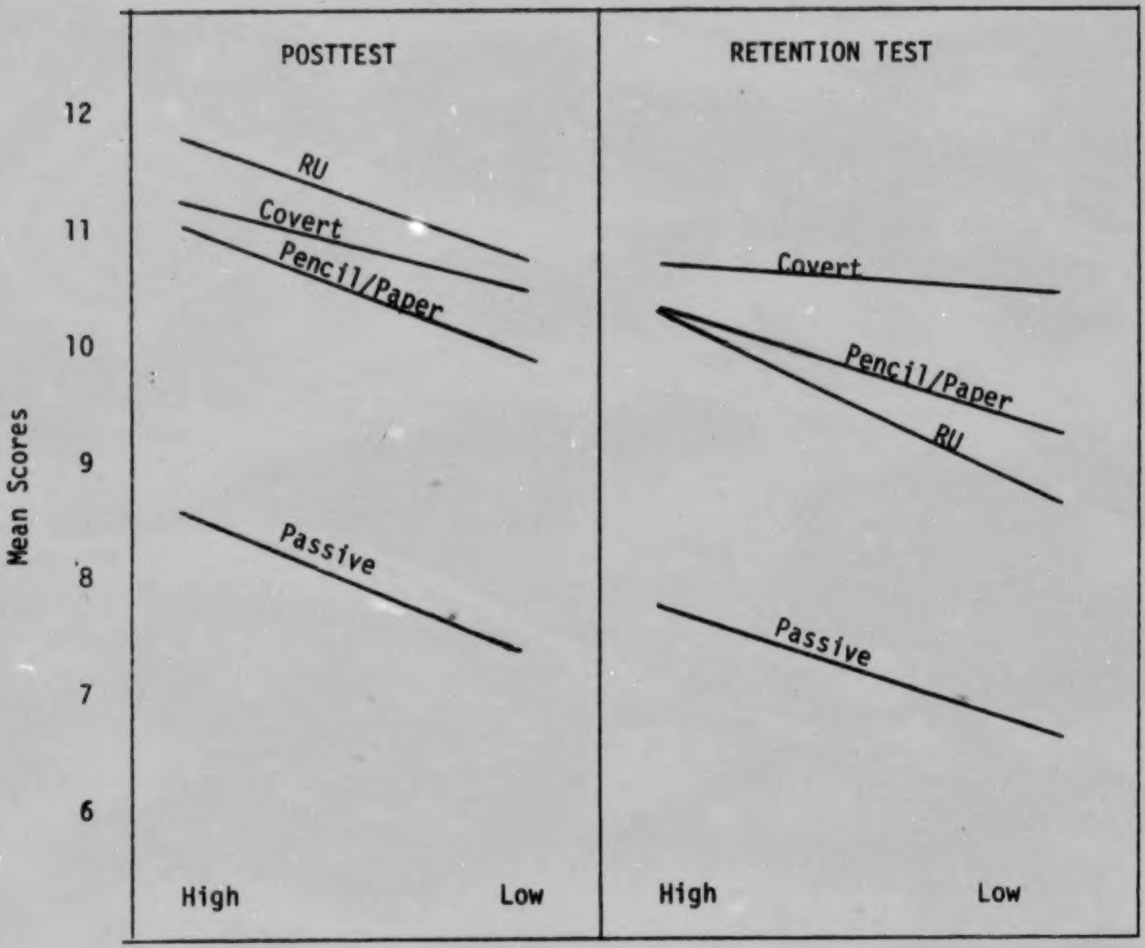
The retention test analysis of variance also resulted in a highly significant difference in methods of instruction at the .0005 level of probability as on the posttest. Achievement levels had a significant difference at the .047 level rather than at the .001 level on the posttest. No interaction variance was found on the retention test; the same as for the posttest. Scheffe's test of means comparison revealed only one pair of means (12 possible) having a significant difference at an acceptable level of probability as shown in Table 4.

Table 4. Significant Retention Test Differences of Achievement Levels and Methods

Level	Methods Compared	Mean Difference	P
High Achievers	(No significant differences)		
Low Achievers	Passive-Covert	-4.100	.001*

*Probability at an acceptable level of significance.

Profiles of the Posttest and Retention Test Data are Presented in Figure 3



Profiles of Posttest and Retention Test Data
Figure 3

DISCUSSION

The data and findings of this study of student response in individual television instruction support the related research on group instruction which has generally found that student response is superior to passive instruction. In this study both high and low achievers in each of the response groups exceeded those instructed by conventional passive instruction on their initial learning. However, the method of student response on initial learning was of little consequence since the covert, pencil/paper or responder unit comparisons did not result any significant differences for high or low achievers.

The retention tests indicated that only the low achievers who "thought" the correct responses to questions during the television instruction were superior to the low achievers who received passive instruction. This would suggest that the marking of a test sheet or pushing a correct button may be a distraction to the learning process. These results would seem to imply that higher initial learning scores resulted in higher retention scores but not significantly so except for the covert treatment. This research adds weight to the concept that retention seems to be influenced more by the initial level of learning than by the method.

FENTON W. BRASHEAR
Educational Specialist
U. S. Army Ordnance Center
and School

The author gratefully acknowledges the advice and counsel of Mr. Harold L. Oliver, Educational Advisor, U. S. Army Ordnance Center and School, Aberdeen Proving Ground, Maryland, on developing the experimental design for this study.

SECTION VIII

GENERAL SESSION

PUBLIC POLICY CONCERNING THE USE OF TESTS

by
Raymond L. Lewiski
Education Specialist

Extension Course Institute
Air University

ABSTRACT

The article identifies major events that have impacted on public policy toward testing during the last twenty years. Events were classified into three catagories and described. Concerns that prompted each event were identified and an interpretation of the interrelationship of events was made. Events were classified as stemming from writings of individual authors, publications of professional organizations, and actions taken by the Legislative, Executive and Judicial branches of government.

The publications of the professional organizations and actions taken by the three branches of government were described in detail. The writings of individual writers were acknowledged but not described.

The major concerns of the initiators of events were identified as follows:

a. Problems generated by the use of tests for making employment decisions.

b. Problems of interpretation generated by semantic differences between the psychological and legal disciplines.

c. Problems associated with and resulting from the impact of accountability on the test developer and user.

A noticable trend was the shifting of the foci of prime concerns from an emphasis on professional competency to an increasing emphasis on professional integrity. This in turn resulted in an increasing resortment to a legalism that holds the test developer and user accountable for their respective product and use.

INTRODUCTION

Public policy concerning the use of tests has been shaped by a multitude of events. Notable events occurring within the past twenty years were reviewed and classified. Events were classified on the basis of who or what institution was instrumental in effecting their occurence. It was found that the most important policy shaping events stemmed from three sources: (1) individual writers, (2) professional organizations, and (3) branches of government.

Among the noteworthy writers in the first classification were Whyte 1956, Baritz 1960, Black 1962, Gross 1962, Hoffman 1962, Fincher 1964 and 1973, Enneis 1964-1969, Ash 1966, Krug 1966, and Lopez 1966. Three noteworthy professional organizations were the American Psychological Association (APA), the American Educational Research Association AERA), and the National Council on Measurement in Education (NCME). Noteworthy events from the third classification consisted of executive orders, legislative acts, and court rulings. Tables 1, 2, and 3 chronologically list noteworthy events classified as originating from individual writers, professional organizations, and branches of government.

Events such as noteworthy publications by individual writers were acknowledged in Table I but not described in detail. Noteworthy events that stemmed from professional organizations and branches of government were given more consideration. Each event classified in these two groupings was described, and an interpretation of its relationship to other events was given. Also an attempt was made to identify major trends, issues, and concerns for each event. The first of these events was the publication of "Ethical Standards of Psychologists" by the American Psychological Association.

Ethical Standards of Psychologists

In 1953 the American Psychological Association (APA) published its "Ethical Standards of Psychologists." This publication was one of the first attempts by an association to initiate a formal set of standards governing the use of tests. During the next two decades, a number of other noteworthy publications that pertained to tests were published. Among these were "Technical Recommendations for Psychological Tests and Diagnostic Techniques" in 1954, "Technical Recommendations for Achievement Tests" in 1955, "Standards for Educational and Psychological Tests and Manuals" in 1966, and "Standards for Educational and Psychological Tests" in 1974. In each of these publications, the dominant theme was interpreted as representing the issues and concerns of the time.

One of the major concerns of the association during the early fifties was the prevention test abuses. It was thought that prevention of test abuses could be accomplished by insuring that test constructors and users were properly trained. The prevailing attitude was that professionally developed tests administered by competent persons would keep test abuses at a minimum. Little emphasis was placed on issues concerning the professional integrity of the test constructors or test users: The focus of attention was on their professional competencies. Thus, in keeping with this belief, a major topic in the 1953 publication pertained to the background knowledge required of persons using various types of test.

Tests were classified into three levels: A, B, and C. Examples of tests classified at each level were given and competencies expected of persons who used tests classified at the various levels were described.

Level A tests consisted of those tests that could be administered, scored and interpreted with the aid of a manual by a person who had only a general orientation to the kind of institution or organization for which the test was being administered. Achievement or proficiency tests were examples of tests at this level.

Level B tests required that the tester have some technical knowledge of test construction and use. The tester was required to have competency in supporting psychological and educational fields such as statistics, individual differences, psychology of adjustment, personnel psychology and guidance. Examples of level B tests would include aptitude tests and adjustment inventories that apply to normal population.

Level C tests consisted of tests and aids, and required that the tester have a substantial understanding of testing and supporting psychological fields, together with supervised experience in the use of these devices. Projective tests and individual mental tests would be included at this level.

The major objectives of the 1953 publication were the classification of tests into levels and the establishment of minimum training standards. These objectives were in keeping with the belief that professionally developed tests administered by competent persons would keep tests abuses at a minimum. The possibility of deliberate test abuses and the legal implications that would result from such abuses was not a major concern at this time. The prevailing belief was that test abuses resulted more from ignorance than intent.

Perhaps the greates influence on public policy toward testing resulting from the "Ethical Standards of Psychologists" was its role as a ferment of discussion. The need for a set of guidelines governing the use of tests was recognized by a number of professional associations. Among these were the American Education Research Association and The National Council on Measurements used in Education. The APA publication fermented discussion within these associations and was the beginning of a joint effort between associations to establish guidelines governing testing.

Standard for Educational and Psychological Tests and Manuals

The "Standards for Educational and Psychological Tests and Manuals" was a publication which resulted from the efforts of three associations. The sequence of events leading to this publication was as follows: In March 1954, the American Psychological Association (APA) issued its "Technical Recommendations for Psychological Tests and Diagnostic Techniques." In January 1955, the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME) jointly issued their "Technical Recommendations for Achievement Tests." In the latter part of 1963 a joint APA-AERA-NCME committee of eight members was constituted for the purpose of formulating a set of technical standards accepted by each association. Three years later in 1966 the associations agreed on and jointly published their "Standards for Educational and Psychological Tests and Manuals." This publication was a major contribution toward the formation of a set of standards and recommendations for persons who use educational and psychological measurements.

The members of the three associations who were the architects of the standards did not restrict their efforts solely to the issues of testing technology; they were also concerned about the human aspects of testing. The authors of the 1966 standards described the impact of tests on the individuals tested as follows:

> Psychological and Educational tests are used in arriving at decisions which may have great influence on the ultimate welfare of the persons tested, on educational points of view and practices, and on development and utilization of human resources. Test users, therefore, need to apply high standards of professional judgement in selecting and interpreting tests, and test producers are under obligation to produce tests which can be of the greatest service. The test producer, in particular, has the task of providing sufficient information about each test so that the users will know what reliance can safely be placed on it.[1]

This quotation depicts the attitude that must have prevailed among the developers of the 1966 Standards. The concern about what reliance can safely be placed on a test was of particular note because this concern was to increase in intensity and eventually generate a host of moral and legal issues about test validity. Basically the foci of concerns had shifted from an emphasis on the test constructor's and user's competency to an emphasis on what reliance can safely be placed on a test.

Six topics pertaining to tests were covered in the 1966 Standards: (1) Dissemination of information, (2) Interpreting, (3) Validity, (4) Reliability, (5) Administration and scoring, and (6) Scales and norms. Of these topics, the one that generated the most controversy was the one on validity. The coverage given to validity was criticized as being both imprecise and insufficient. The decriptions of the various types of validity given in the 1966 Standards may have been precise and sufficient for a person schooled in psychometrics but for other readers they were vague and incomplete. Three types of validity were described; namely: Content Validity, criterion-related validity, and construct validity. Future guidelines from the Equal Opportunity Commission were to recommend each of these types of validity for meeting minimum standards for test validation.

The Motorola Case

The case of Motorola, Inc. Vs. Illinois Fair Employment Practices Commission and Leon Myart was the first major court case pertaining to the use of tests for employment purposes. The main events of this case beginning 15 July 1963 and ending 24 March 1966 were outlined by Wallace, Kissinger and Reynolds in Appendix A of the article "Testing of Minority Group Applicants for Employment." The following was based on their chronology of the Motorola Case.

On 15 July 1963 Leon Myart, a black, applied for a job as a television phaser and analyzer at the Franklin Plant of Motorola, Inc., in Franklin, Illinois. He was required to take a five-minute intellignece test administered by the company and was interviewed. He was not offered a job, nor was he informed whether he qualified for the job.

On 29 July 1963 Myart filed a complaint with the Illinois Fair Employment Practices Commission and the President's Committee on Equal Employment Opportunity alleging that his not being hired was due to racial discrimination. The case was examined by Robert Bryant of the Illinois Fair Employment Practices Commission on 27 and 28 January 1964, and the question whether a general ability test could be used as a basis for selection for employment was raised. Upon examining the case, Bryant ruled that the intellignece test, General Ability Test no. 10, could not be used as a basis for disqualifying Myart for employment because the test had been normed on "advantage groups" and did not "lend itself to equal employment opportunity to qualify for the hitherto culturally deprived and disadvantaged groups." He therefore directed Motorola to offer Myart a job and to discontinue using test no. 10.

The decision was appealed and the case was revieded before the full commission four times during 1964. Reviews of the case were held on April 18, May 25, July 14-15, and November 18.

On November 18 a unanimous decision was issued by the commission. The members of the commission agreed that Myart had been denied employment because of his race, but they did not direct that he be hired by the company. They did direct however, that he be compensated one-thousand dollars by Motorola, Inc.

The case was appealed to the Illinois Circuit Court. The circuit court reversed the decision that Myart be paid one-thousand dollars compensation; however, the findings on discrimination were upheld.

The case was then appealed to the Illinois Supreme Court. On March 24, 1966, the Illinois Supreme Court reversed the judgement of the circuit court on the grounds that the alleged unfair employment practice was not established by a preponderance of evidence.

This ended a case that had been under review for over two and one-half years, a case that had generated controversy and undergone a number of reverse decisions, a case that had received national publicity and, finally, a case that was a landmark in judicial involvement in public policy toward testing. The importance of the case was not attributed to the final ruling of the Illinois Supreme Court. The importance of the case was due to the publicity that was given to the directives issued by the Illinois Fair Employment Commission. The directives issued by the commission on January 1964 gained national attention. The impact of the publicity was in turn felt in congress where civil rights legislation was being debated. This in turn prompted Senator Tower, a Republican from Texas, to attach an amendment to the Civil Rights Act of 1964.

The Tower Amendment to The Civil Rights Act of 1964

During 1964, civil rights legislation was being debated in Congress. The topics of debate covered a wide spectrum, including public education, public facilities, housing, employment practices, etc. The question of unlawful employment practices was raised and debated. During debate of this issue, an amendment was introduced by Senator Tower of Texas. This amendment later became the Tower Amendment to Title VII of the Civil Rights Act of 1964.

The Tower Amendment was prompted by the Illinois Fair Employment Practice Commission's ruling on the Motorola Case. The commission had disbarred the use of general ability tests by employers, if said tests were to be used as a basis for selecting persons to be hired. The Tower Amendment had the effect of establishing, in legal terms, the right of an employer to use "professionally developed ability tests" for the purpose of selecting persons for employment provided that the tests were not intentionally discriminatory. Thereby, the amendment represented a major attempt by the legislative branch of government to affect public policy governing the use of tests.

The Tower Amendment was embodied in Section 703 (h) of Title VII of the Civil Rights Acts of 1964. It reads as follows:

> Nor shall it be an unlawful employment practice for an employer to give and act upon the results of any professionally developed ability test provided that such tests, its administration or action upon the results is not designed, intended or used to discriminate because of race, color, religion, sex, or national orgin.[2]

The wording of the amendment led to a controversy over the use of and meaning of the term "ability test". The question was raised whether a legal interpretation of the term "ability test" was to include broad general abilities that tend to measure the person in the abstract or be restricted to specific job related abilities that could be empirically verified. In other words, were mental ability tests or intellignece tests also to be classed as "ability tests?" Members from both the psychological and legal professions were in disagreement over how the term "ability test" was to be interpreted. An answer to this question was not easily forthcoming.

The Emergence of Executive Influence on Public Policy Toward Testing
Executive Order No. 11246

On 24 September 1965 President Johnson issued Executive Order 11246 entitled "Equal Employment Opportunity". This order superseded Executive Order No. 10590 (19 June 1955), 10722 (5 August 1957), 10925 (6 March 1961), 11114 (22 June 1963) and 11162 (28 June 1964).

The order consisted of four parts, three of which indirectly pertained to public policy toward the use of tests. These three were (1) nondiscrimination in government employment, (2) nondiscrimination in employment by government contractors and subcontractors, and (3) nondiscrimination provisions in federally assisted construction contracts.

Part I consisted of policy statements by the executive branch of government on discrimination in federal employment. In general, a policy of equal opportunity was to apply to every aspect of federal employment. Discrimination on the basis of race, creed, color, or national orgin was specifically prohibited. Each executive department and agency was required to use an existing staff or establish an office for the purpose of carrying out these policies. For example, the task of insuring equal employment opportunity programs for civilian employees was assigned to the existing staff of the Civil Service Commission. Other governmental departments, such as the Department of Labor, had to establish an office to carry out such parallel functions.

Parts II and III dealt with discrimination as it related to employment practices by government contractors, subcontractors and federally assigned construction contractors. The topics covered in Part II were (A) duties of the Secretary of Labor, (B) contractors' agreements, (C) powers and duties of the Secretary of Labor and the contracting agencies, (D) sanctions and penalties, and (E) certificates of merit; whereas, Part III dealt with nondiscrimination provisions in federally assisted construction contracts.

The issuance of Executive Order 11246 set the stage for an eventual impact on public policy toward testing. One feature of the order was a requirement that nondiscrimination clauses be included in all government contracts, sub-contracts, and federally assisted construction contracts. The Secretary of Labor was charged with the responsibility for insuring that contractors comply with such clauses. To carry out this obligation, the Secretary of Labor established The Office of Federal Contract Compliance, Equal Employment Opportunity, Department of Labor. In turn, authority was delegated to this office for carrying out its responsibilities. One responsibility of this office was to develop and issue guidelines governing employment practices. Thus, authority had been delegated and the stage set for an eventual impact by an agency of the executive branch of government on public policy toward the use of tests.

The Executive Branch of Government Acknowledges Sex as a Possible Basis for Discrimination (Executive Order 11375)

A noticeable phenomenon about the text of Executive Order 11246 was the exclusion of sex as a possible basis for discrimination. The exclusion of sex was particularly noticeable in view of the fact that Title VII of the Civil Rights Act of 1964 dealt with the problem of Equal Employment Opportunity, and Section 703 of Title VII was entitled "Discrimination because of Race, Color, Religion, Sex or National Orgin;" whereas, Executive Order 11246 entitled "Equal Employment Opportunity" issued almost a year after the passage of the 1964 Civil Rights Act excluded sex as a possible basis for discrimination. Order 11246 acknowledged race, creed, color and national orgin as possible basis for discrimination but not sex. This remained the official position of the executive branch of government until 1967 when Executive Order 11375 was issued. In effect, Order 11375 amended 11246 to include sex as a possible basis for discrimination. In each case "Discrimination because of race, creed, or national orgin" was amended to read "Discrimination because of race, color, religion, sex or national orgin."

The Equal Employment Opportunity Commission's "Guidelines on Equal Employment Testing Procedures"

On 2 July 1964 civil rights legislation was passed by Congress that would eventually have a considerable effect on public policy toward the use of tests. A year later on 2 July 1965 the act became effective. One intent of the act was to insure that all Americans would be considered for hiring and promotion on the basis of their ability and qualifications, without regard to race, color, religion, sex or national orgin. To insure that this intention would be carried out, the act provided for the establishment of the Equal Employment Opportunity Commission (EEOC).

A prime responsibility of the EEOC was to conduct technical studies in support of the purposes of Title VII. It soon became apparent that the purposes of Title VII could be interpreted in more than one way; of particular note were the different interpretations given for Section 703 (h), the

Tower Amendment. A major dispute developed over how this section was to be interpreted. To resolve this dispute, the EEOC formulated a set of guidelines which were published in the Federal Register on 24 August 1966. The publication was entitled "Guidelines on Equal Employment Testing Procedures." This publication was the first set of guidelines to be issued by the EEOC, and they were to be met with both success and failure.

From its very beginning, the EEOC was embroiled in controversies--some semantical, others attitudinal. From the semantical side, the wording of the Tower Amendment had led to controversy over what constituted an ability test, while on the attitudinal side there was a noticeable shift of the foci of prime concerns from an emphasis on professional competency to an emphasis on professional integrity.

The semantical problem has been viewed as resulting from a communications gap between two distinct disciplines, namely those of psychometrics and law. In 1969 Enneis perceived the problem as follows:

> During the four years that the Equal Employment Opportunity Commission has been in operation, the issues of employment testing have been among the more persistent and difficult ones. As far as the EEOC is concerned, it is probably correct to say that psychological testing involves the most direct confrontation of a scientific discipline with legal definitions of employment discrimination.[3]

The resolution of semantic differences between the psychological and legal disciplines was to be the major concern of forthcoming guidelines by the EEOC.

From the standpoint of attitude, the shift in concerns meant that issues would no longer focus merely on the competency or incompetency of the test constructor and user; it would also focus on their intentions and professional integrity. The following quote from the 1966 EEOC guidelines exemplifies both the emerging semantical issues and shifting of concerns.

> The commission accordingly interprets professionally developed ability tests to mean a test which fairly measures the knowledge or skill required by the particular job or class of jobs which the applicant seeks, or which fairly offers the employer a chance to measure the applicant's ability to perform a particular job or class of jobs. The fact that a test was prepared by an individual or organization claiming expertise in test

preparation does not, without more, justify its use within the meaning of Title VII.[4]

In brief, the 1966 guidelines were vaguely written and tended to raise issues, not resolve them. The replacement of "professionally developed ability tests" by "a test which fairly measures knowledge or skills and fairly offers the employer a chance to measure the applicant's ability" did little toward resolving the controversy. The guidelines did not indicate what a fair measure would be or how fairness would be judged. Also, what constituted "more" in "without more" in the above quotation was not clearly defined; at best, it alluded to an emerging attitude that test developers and users ought to be accountable for their respective products and uses.

Despite the acknowledged shortcomings of the 1966 EEOC Guidelines, their publication was a noteworthy event. They were the first set of guidelines to be issued in compliance with the 1964 Civil Rights Act and represented a major influence, stemming from the legislative branch of government, on public policy toward testing. They were to remain in effect until 1 August 1970, at which time they would be superseded by a new set of guidelines entitled "Guidelines on Employee Selection Procedures."

The Equal Employment Opportunity Commissions "Guidelines on Employee Selection Procedures"

On the first of August 1970 a new set of guidelines entitled "Guidelines on Employee Selection Procedures" was issued by the EEOC. The new guidelines superseded and enlarged upon the 1966 guidelines. A noticeable trend in the new guideline was the tendency to describe in more detail the topics discussed and terminology used. Two terms of particular importance were "test" and "discrimination."

The term "test" was defined as any paper and pencil or performance measure used as a basis for any employment decision. Paper and pencil or performance measures included measures of general intelligence; mental ability and learning ability; specific intellectual abilities; mechanical, clerical and other aptitudes; dexterity and coordination; knowledge and proficiency; occupational and other interests; and attitudes, personality or temperament. Also included were all formal, scored quantified or standardized techniques of assessing job suitability such as specific qualifying or disqualifying personal history or background requirements, specific educational or work history requirements, scored interviews, biographical information blanks, interviews rating scales, scored application forms, etc. Any of the above measures, when used as a basis for making employment decisions, constituted a "test", whereas employment decisions included decisions affecting one's eligibility for hire, transfer, promotion, membership, training, referral or retention.

Discrimination was defined as the use of any test which adversely affects hiring, promotion, transfer or any other employment or membership opportunity of classes protected by Title VII of the Civil Rights Act of 1964 unless: (a) the test has been validated and evidences a high degree of utility and (b) the person giving or acting upon the results of the particular test can demonstrate that alternative suitable hiring, transfer or promotion procedures are unavailable for his use.

Topics on validity were discussed in detail. Among these were topics on Evidence of Validity, Minimum Standards for Validation, Presentation of Validity Evidence, Use of Other Validity Studies and Assumption of Validity.

The guidelines disallowed any assumption of validity and placed the burden of proof of validity on the tester. Under no circumstances could the general reputation of a test, its author or its publisher, or casual reports of a test's utility be accepted in lieu of evidence of validity. In every case in which a test was used for making an employment decision, evidence of the test's validity was to be supplied by the tester.

Validation requirements varied with each testing situation. Under certain conditions, content or construct validity were acceptable for meeting validity requirements. For example, evidence of content validity alone would be acceptable provided that the test were well developed and consisted of suitable samples of the essential knowledge, skills or behaviors composing the job in question. However, when technically feasible, a criterion-related validity was required and could not be substituted for by content or construct validities. The state of being "technically feasible" was described in the guidelines as follows:

> The term "technically feasible" as used in these guidelines means having or obtaining a sufficient number of minority individuals to achieve findings of statistical and practical significance, the opportunity to obtain unbiased job performance criteria, etc.[5]

Persons claiming the absence of technical feasibility were required to demonstrate positive evidence of this absence. In other words, evidence of content (in the case of job knowledge or proficiency tests) or construct (in the case of trait measures) validity, would be acceptable only if it could be positively demonstrated that a criterion-related validity was technically unfeasible.

The "Standards for Educational and Psychological Tests and Manuals" published by the American Psychological Association (APA) were specifically referred to in the topic on "Minimum Standards for Validation." Evidence of content, construct or criterion-related validity called for in the guidelines was to be subjected to the same standards for those types of validity as described in the APA publication. For example, if content validity were allowed as evidence of test validity, then the standards governing content validity in the APA publication would be applicable to the tester. This included such requirements as obtaining an adequate sample, describing the credentials of the experts who selected the items, and reporting the extent of agreement between independent judgements about each item.

Another type of validity referred to in the guidelines but not mentioned in the APA publications was differential validity. Differential validity was included as a validity requirement in the 1970 guidelines. A precise definition of differential validity was not offered; however, various aspects that pertained to the notion of differential validity were described. In brief, to meet the differential validity requirements, data must be generated and results separately reported for minority and non minority groups whenever technically feasible. Here again the question of technical feasibility was raised and the burden of demonstrating technical unfeasibility fell on the tester.

The 1970 guidelines, like the 1966 guidelines, clarified some questions and generated others. Detailed descriptions of terms such as "test" and "discrimination" tended toward clarification; however, vaguely defined phrases such as "differential validity" and "technically feasible" insured that semantic differences between the legal and psychological disciplines would remain.

Despite these shortcomings in the intended goal of clarifying meaning, the 1970 guidelines met with some success and marked a high point in legislative influence on public policy toward testing. The use of tests for employment decisions had increased considerably since the passage of the 1964 Civil Rights Act, and the number of possible misuses of tests had likewise increased. Thus, the 1970 guidelines provided a much needed interpretation of the intent of Title VII of the 1964 act and, thereby, contributed to the implementation of nondiscriminatory personnel policies. Then too, in all likelihood these guidelines abetted the ferment for a court test of Title VII. Such a court test was soon to be forthcoming in Griggs Vs. Duke Power Company.

Griggs Vs. Duke Power Company

The case of Willie S. Griggs et al., Petitioners., Vs. Duke Power Company was the second major court case pertaining to public policy toward the use of tests. The first case, Myart Vs. Motorola, was decided in the Illinois Supreme Court on March 1966. The Griggs Vs. Duke Power Company case was decided in the U.S. Supreme Court on the 8th of March 1971. The former case differed from the latter case in both the level of the deciding court and the direction of the decision rendered. Unlike the former case, the Griggs Vs. Duke Power Company case was a U.S. Supreme Court decision and in favor of the original petitioners. Consequently it has had a greater effect on public policy toward the use of tests than the Motorola case.

The Griggs Vs. Duke Power Company case began as a class action suit by black employees at Duke Power Company's Dan River steam station located at Draper, North Carolina. The station had 95 employees, 14 of whom were black. Thirteen of the 14 black employees were petitioners in the suit. The plantiffs claimed that the company's employment practices violated their civil rights. Specifically challedged was the company's employment requirement that any employee or potential employee have a high school education or pass a standardized general intelligence test as a condition for transfer to a new job (in the case of an old employee) or initial employment (in the case of a new employee). The petitioners claimed that these requirements for hire and transfer operated to render ineligible a disproportionate number of blacks.

The case was brought before the United States District Court for the middle district of North Carolina at Greensboro, North Carolina. The District Court dismissed the complaint. The dismissal was appealed and the appeals court remanded that in absence of a discriminatory purpose, requirement of a high school education or the passing of a standardized general intelligence test as a condition of employment in or transfer to jobs was permitted by the Civil Rights Act. The claim was rejected that such requirements operated to render ineligible a markedly disproportionate number of Negroes and were therefore unlawful under the act unless shown to be job-related. Certiorari was granted and the case was sent to the Supreme Court for review. The opinion of the court was delivered by Chief Justice Burger. The Supreme Court reversed the appeals court's remand, holding that the employer was prohibited by provisions of the civil rights act from requiring a high school education or passing of standardized general intelligence where (a) neither standard was shown to be significantly related to successful job performance, (b) both requirements operated to disqualify Negroes at a substantially higher rate than white applicants, and (c) jobs in question formerly had been filled only by white employees as part of a longstanding practice of preference to whites. This decision was to have a profound impact on public policy toward testing. Prior to the Supreme Court's ruling, the EEOC's "Guidelines on Employee Selection Procedure" offered an interpretation of the intent of the Civil Rights Act of 1964 and recommended necessary measures an employer should take to be in compliance with the act's intent. By contrast, the Supreme Court's decision was a definitive statement of the act's intent and, henceforth, tests used for the purpose of making employment decisions were required by law to be job-related, and the burden of proving job-relatedness was placed on the employer.

Justice Burger interpreted Congressional intent as follows:

> Nothing in the act precludes the use of testing or measuring procedures; obviously they are useful. What Congress has forbidden is giving these devices and mechanisms controlling force unless they are demonstrably a reasonable measure of job performance. Congress has not recommended that the less qualified be preferred over better qualified simply because of minority orgins. Far from disparaging job qualifications as such, Congress has made such qualifications the controlling factor, so that race, religion, nationality, and sex become irrelevant. What Congress has commanded is that any tests used must measure the person for the job and not the person in the abstract.[6]

Basically the foci of concerns had shifted from emphasis on the test constructor's and user's competency to an emphasis on what reliance could safely be placed on a test.

The Supreme Court's decision rendered a definitive statement governing some aspects of testing and employment decision making; however, other aspects were not considered in this case. For example, no ruling was rendered about the legal status of testing requirements that take into account capabilities for the next succeeding position. For instance, in a context in which a potential for future promotion could be shown to be a genuine business need, a test that was shown to be job-related to succeeding positions might be considered as a bonified test. At this time the status of such a test has not been determined.

Part 60-3 -- Employee Testing and Other Selection Procedures

Less than two months after the Griggs Vs. Duke Power Company case was decided, a notice of proposed rule making for amending Chapter 60 of Title 41 was published in the Federal Register. The usual 30 days were given to interested persons to submit written comments, suggestions, or objections regarding the proposed amendments and finally, on 2nd October 1971, an amendment to chapter 60, "Part 60-3 Employee Testing and Other Selection Procedures" was published in the Federal Register and became effective on that day.

The guidelines given in "Part 60-3 Employee Testing and Other Selection Procedures" were basically the same as the "Guidelines on Employee Selection Procedures" issued on the 1st of August 1970 by the EEOC. Both sets of guidelines were intended to impose the same basic requirements on employers, contractors, or other persons covered by them. The guidelines differed only in the language arising from the different legal authority of the two agencies. This was noteworthy in that it was one of the first deliberate attempts to arrive at a concensus between governmental agencies toward the question of what should be contained in their guidelines. The question of what should be included in a uniform set of guidelines applicable to all governmental agencies was to become a major concern for the EEOC.

The Equal Employment Opportunity Act of 1972

On March 24, 1972, The Equal Employment Opportunity Act (Public Law 92-261) was approved. The enactment of this law was a major step forward in assuring an equal employment opportunity to minority groups and women in our society. In turn, it was a major event affecting public policy toward the use of tests. The '72 act amended Title VII of the Civil Rights Act of 1964, expanding its coverage and adding enforcement powers.

The duties and responsibilities of the EEOC were broadened, giving it quasi-legal enforcement powers. Section 706 (a) as amended empowered the commission to prevent any person from engaging in any unlawful employment practices as set forth in Section 703 (Discrimination Because of Race, Color, Religion, Sex, or National Orgin) and Section 704 (Other Unlawful Employment Practices). Cases involving a government, governmental agency, or political subdivision were charged to the Attorney General. Section 707 (c) as amended provided for the commission's assumption of the functions charged to the Attorney General and his office. These functions were to be assumed within two years after the date of enactment of the Equal Employment Opportunity Act of 1972.

Another amendment that will most likely have a consierable effect on public policy toward the use of tests was Sec. 715. This section provided for the establishment of the Equal Employment Opportunity Coordination Council (EEOCC) and reads as follows:

> There shall be established an Equal Employment Opportunity Coordinating Council (hereinafter referred to in this section as the Council) composed of the Secretary of Labor, the Chairman of the Equal Employment Opportunity Commission, the Attorney General, the Chairman of the United States Civil Service Commission, and the Chairman of the United States Civil Rights Commission, or their respective delegates. The Council shall have the responsibility for developing and implementing agreements, policies and practices designed to maximize effort, promote efficiency, and eliminate conflict, competition, duplication and inconsistency among the operations, functions and jurisdictions of the various departments, agencies and branches of the Federal government responsible for the implementation and enforcement of equal employment opportunity legislation, orders, and policies. On or before July 1 of each year, the Council shall transmit to the President and the Congress a report of its activities, together with such recommendations for legislative or administrative changes as it concludes are desirable to further promote the purposes of this section.[7]

The Council has issued a discussion draft of its pending "Uniform Guidelines on Employee Selection Procedures." These should be approved sometime this year (1974).

Standards for Educational and Psychological Tests

The "Standards for Educational and Psychological Tests" was published in 1974 by the American Psychological Association (APA). The publication was the joint work of many psychologists and educators. The principal author was Professor Robert M. Guion, Bowling Green State University, Bowling Green, Ohio. Basically, the publication was a revision of the 1966 "Standards for Educational and Psychological Tests and Manuals." Like the 1966 publication, the 1974 publication was a joint effort by the American Psychological Association, the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME).

The publication of the 1974 Standards was prompted by a number of events. First, the guidelines issued by the Equal Employment Opportunity Commission in August 1970 made reference to the 1966 Standards. This focused attention on the standards which were lacking in clarity on some topics and did not adequately cover others. Then on 8 March 1971 the historic Supreme Court decision of Griggs Vs. Duke Power Company impacted further on public policy toward the use of tests. This decision greatly increased the number of discrimination suits and in turn increased the need for a new set of standards. Shortly thereafter, on 2 October 1971, the Office of Federal Contract Compliance, Equal Employment Opportunity Department of Labor, issued a set of guidelines by authority of Executive Order 11246. Again specific reference was made to the 1966 APA Standards. By this time it was apparent that the 1966 Standards needed to be revised. Finally, as plans for revision were underway the Equal Employment Opportunity Act of 1972, Public Law (92-261) further influenced public policy toward the use of tests. The 1972 Act broadened the coverage of Title VII of the 1964 Act and included employment decisions by educational institiutions within its jurisdiction. Each of these events had its impact on the 1974 Standards.

Another factor impacting on the 1974 Standards was the prime concerns of its authors. The prime concerns of the authors of the 1974 Standards differed somewhat from the concerns expressed by the authors of the 1966 Standards. The authors of both publications shared similar concerns about the human aspect of testing. However, changing conditions dictated some new concerns. The following quote from the 1971 Standards reflects some of these concerns:

> Part of the stimulus for revision is awakened concern about problems like invasion of privacy of discrination against members of groups such as minorities or women. Serious misuses of tests include, for example, labeling Spanish-speaking children as mentally retarded on the basis of scores on tests standardized on a "representative sample of American children," or using a test with a major loading on verbal comprehension without appropriate validation in an attempt to screen out large numbers of blacks from manipulative jobs requiring minimal verbal communication.[8]

The examples cited above were among the concerns shared by the authors of 1974 Standards.

Another factor impacting on the 1974 Standards was an emerging notion of accountability. Accountability has become a concern for most professional groups; ie, teachers, administrators, politicians, etc. Accountability as it applies to public policy toward testing requires that both the test constructor and test user be held responsible for their respective product or use. The requirement that they be responsible raises some complex questions, and poses problems that have both ethical and legal ramifications. Questions such as "how would responsibility be enforced" and "what conditions were necessary before responsibility could be actualized" were among those raised. The mere delegation of responsibility does not insure accountability. To actualize accountability, two things must be done. First, some person or agency must scrutinize the test developer and tester and second, the prerequisite conditions to responsibility must be identified and monitored. Two such prerequisites to responsibility have been identified. These were (1) that the test developer and tester be competent and (2) that they be fair. The inclusion of fairness as a prerequisite to responsibility meant that the scrutiny of the individual in question would not be limited solely to his professional competency but would also be directed toward his professional integrity. This represented a major shift from the previously held belief that all that was necessary to prevent test abuses was to insure that persons involved in testing were competent.

The authors of the 1974 Standards were sensitive to the concerns and issues that resulted from using tests for making employment decisions. This was attested to by the numerous examples pertaining to employment practices cited in the publication. They were also sensitive to the semantic differences that existed between the psychological and legal disciplines. This was attested to by the coverage given to the topic on validity.

Validity has been a particularly troublesome concept. The topic on validity in the 1966 Standards was lacking in both coverage and clarity. From a psychometric point of view the 1966 descriptions may have been acceptable, but from a legal point of view they were inadequate. In view of the deficiency, a greater emphasis was placed on the topic of validity in the 1974 Standards and examples cited for content, construct, criterion related, and differential validities were more definitive. It is important to note that the standards were only recommendations and had no legal status per se; however, the issues reviewed pertaining to enforcement, competency and fairness were also vital legal issues. Thus, the 1974 Standards will most likely influence forthcoming legal opinions and may resolve many of the semantic differences that exist between the psychological and legal disciplines.

Some issues have not been resolved. For example, the relationship of validity to the notion of job relatedness has been intrepreted differently by different persons. Guion[9] suggested that an employment test may provide a basis for inferences that have criterion-related validity, or construct validity, or content validity, or all of these, and still not be job related, Guion was unwilling to equate "validity" and "job relatedness." Also, there is no general agreement on the prior question of what is fairness and how it can be measured. Various math models have been suggested for assessing fairness. Cleary[10] (1968) suggested that a regression analysis be used as a basis for determining and measuring fairness. Thorndike[11] (1971) recommended that a quota system be used and Darlington[12] (1971) approached the problem in terms of a partial r, correlational technique. As of now the issues surrounding the notion of job relatedness and fairness have not been resolved. Possible, these issues will be clarified in the forthcoming "Uniform Guidelines on Employer Selection Procedures."

TABLE I

A CHRONOLOGICAL LIST OF ARTICLES THAT HAVE HAD A NOTEWORTHY EFFECT ON PUBLIC POLICY TOWARD THE USE OF TESTS

Date	Author	Article
1956	Whyte, W. H., Jr.	The Organization Man. Garden City, N.Y., Doubleday , 1956
1960	Baritz, L.,	The Servants of Power, Middletown, Conn,; Wesleyan University Press, 1960.
1962	Gross, M. L.,	The Brain Watchers. New York; Random House, 1962.
1962	Hoffman, B.,	The Tyranny of Testing. New York, Crowell-Collier, 1962.
1963	Black, H.,	They Shall Not Pass. New York,Morrow; 1963.
1964	Fincher, C.,	Testing Critics and Criticism. Atlanta Economic Review, 1964, 3-7.
1966	Ash, Philip,	The Implications of the Civil Rights Act of 1964 for Psychological Assessment in Industry. American Psychologist, 1966, 21, 797-703.
1966	Krug, R. E.,	Some Suggested Approaches for Test Development and Measurement. Personnel Psychology, 19, No. 1, 24-35.
1966	Lopez, Felix, Jr.,	Current Problems in Test Performance of Job Applicants. Personnel Psychology, 19, No. 1, 10-18.
1967	Enneis, William H.,	Statement Before House Post Office and Civil Service Subcommittee.*
1967	" " "	Discrimination: Planned and Accidental.*
1969	" " "	Personnel Testing and Equal Employment Opportunity.*
1969	" " "	Misuses of Tests.*
1969	" " "	Minority Employment Barriers from the EEOC Viewpoint.*
1970	" " "	Statement on Personnel Testing and Selection.*
1970	" " "	Use of Nontest Variables in the Government Employment Setting.*

* Equal Employment Opportunity Commission, Personnel Testing and Equal Employment Opportunity, U.S. Government Printing Office, Washington, D.C. 20402, Dec, 1970.

TABLE II

A CHRONOLOGICAL LIST OF NOTEWORTHY EVENTS INITIATED BY PROFESSIONAL ORGANIZATIONS THAT HAVE AFFECTED PUBLIC POLICY TOWARD TESTING

	EVENT	SOURCE	DATE
1.	Publication of "Ethical Standards of Psychologists."	American Psychological Association (APA)	1953
2.	Publication of "Technical Recommendations for Psychological Tests and Diagnostic Techniques."	American Psychological Association (APA)	March 1954
3.	Publication of "Technical Recommendations for Achievement Tests."	American Educational Research Association (AERA) and The National Council on Measurement in Education. (NCME)	Jan 1955
4.	Publication of "Standards for Educational and Psychological Tests and Manuals."	APA, AERA, and NCME	1966
5.	Publication of "Standards For Educational and Psychological Tests."	APA, AERA, and NCME	1974

TABLE III

A CHRONOLOGICAL LIST OF EXECUTIVE, LEGISLATIVE AND JUDICIAL EVENTS THAT HAVE SHAPED PUBLIC POLICY TOWARD TESTING

	EVENT	SOURCE	DATE
1.	Myart vs. Motorola Co.	Judicial	15 July 1963 - 24 March 1966
2.	The Civil Rights Act of 1964	Legislative P.L. 88-352	2 July 1964
3.	Issuance of Executive Order on Equal Employment	Executive Order 11246	28 September 1965
4.	Publication of EEOC's "Guidelines on Equal Employment Testing Procedures"	Legislative 88-352	24 August 1966
5.	Amendment of Executive Order 11246	Executive Order 11375	17 October 1967
6.	Publication of EEOC's "Guidelines on Employee Selection Procedures"	Legislative 88-352	1 August 1970
7.	Willie S. Griggs et al, Petitioners, Vs. Duke Power Company	Judicial 401 U.S. 124	8 March 1971
8.	Publication of "Part 60-3 Employee Testing and other Procedures" by Office of Federal Contract Compliance, Equal Employment Opportunity, Department of Labor	Executive Order 11246	2 October 1971
9.	The Equal Employment Opportunity Act of 1972	Legislative, Public Law (92-261)	March 1972

FOOTNOTES WITH REFERENCES

1. American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (joint committee). Standards for Educational and Psychological Tests and Manuals. Washington, D.C.; American Psychological Association, Inc. 1966. page 1.
2. Public Law 88-352, as approved by the President on July 2, 1964; Civil Rights Act of 1964 with Explanation Commerce clearing House, Inc., Chicago 46, Illinois 1964, page 71.
3. William H. Enneis; Personnel Testing and Equal Employment Opportunity Personnel Testing and Equal Employment Opportunity Commission; U.S. Government Printing Office Washington, D. C. 20402 1970, page 25.
4. Equal Employment Opportunity Commission. Guidelines on Employment Testing Procedures. Washington; EEDC, 1966.
5. Equal Employment Opportunity Commission. part-1607-Guidelines on Employee Selection Procedures; Federal Register Volume 35, Number 149 Saturday, Aug 1, 1970. Washington, D.C. pages 12333-12336.
6. Supreme Court of The United States. Griggs et. al.versus Duke Power Company. No. 124-October Term, 1970 (March 8, 1971).
7. Subcommittee on Labor of the Committee on Labor and Public Welfare United States Senate; The Equal Employment Opportunity Act of 1972; U.S. Government Printing Office; Washington 1972. page 16.
8. American Psychological Association, American Educational Research Association, and National council on Measurement in Education (Joint Committee), Standards for Educational and Psychological Tests; American Psychological Association Inc. 1200 Seventeenth Street, N. W. Washington, D.C. 20036; 1974, page 1.
9. Guion, Robert M. Open a New Window:Validities and Values in Psychological Measurement; American Psychologist Vol. 29, number 5; The American Psychological Association, Inc. May 1974, page 290.
10. Cleary, Anne T. Test bias: Prediction of Grades of Negro and White Students in Integrated Colleges. Journal of Educational Measurement, Vol. 5, No.2, Summer 1968, pages 115-124.
11. Thorndike, Robert L. Concepts of Culture-Fairness, Journal of Educational Measurement Vol. 8, No. 2, Summer 1971. pages 63-70.
12. Darlington, Richard B. Another Look at "Cultural Fairness", Journal of Educational Measurement, Vol. 8, No. 2, Summer 1971, pages 71-82.

OTHER REFERENCES

American Educational Research Association, and the National Council on Measurement in Education. "Technical Recommendations for Achievement Tests" 1955.

American Psychological Association; American Educational Research Association, and National Council on Measurements Used in Education (Joint Committee). Technical Recommendations for Psychological tests and diagnostic techniques. Psychological Bulletin, 1954, 51,201-238.

American Psychological Association."Ethical Standards of Psychologists." Washington, D.C., 1953.

Executive Order 11246. Equal Employment Opportunity, Federal Register, Sept, 28, 1965. 12319.

Executive Order 11375. Amending Executive Order No. 11246, Relating to Equal Employment Opportunity. Federal Register, Oct, 17, 1967, 14303.

Office of Federal Contract Compliance, Equal Employment Opportunity, Department of Labor; Part 60-3. Employee Testing and Other Procedures. Federal Register, 1971, 36 (77), 7532-7535.

DEPARTMENT OF THE ARMY
UNITED STATES ARMY INFANTRY SCHOOL
FORT BENNING, GEORGIA 31905

United States Army Infantry School presentation to the 1974 Military Testing Association Conference, Hilton West Hotel, Oklahoma City, Oklahoma, 21-25 October 1974

THE COMMON BOND IN PEER AND PRECEPTOR RATINGS

Captain John C. Helmick, Jr.

Presented by:
J. E. Gerber, Jr.

THE COMMON BOND IN PEER AND PRECEPTOR RATINGS

--Captain John C. Helmick, Jr.--

ABSTRACT

This paper reports on the development of a combined peer (student associate) and preceptor (tactical officer) rating instrument as a possible indicator for leadership development and evaluation of lieutenants enrolled in the Infantry Officer Basic Course at the U. S. Army Infantry School.

BACKGROUND

All newly commissioned Army lieutenants assigned to the Infantry Branch, as a rule, must complete a standard twelve-week basic officer course. The purpose of this training is to insure that each officer will be able to perform those tasks required by his first duty assignment. Emphasis within the course is placed on leadership and the fundamentals, weapons, equipment and techniques required in order to prepare the graduate for immediate assignment. Practical work is stressed, with theoretical instruction being kept to a minimum. Instruction includes combined arms tactics, military leadership, staff subjects, general military subjects, communications and electronics, unit and material readiness, weapons, and physical fitness training. Particular attention within the course is directed toward the concept of training to train.

INTRODUCTION

Previously within the Infantry Officer Basic Course (IOBC), the principal indicators of an officer's leadership skill were the grades he received on academic tests of leadership subjects taught during the course. Even though these academic tests were performance oriented and criterion-referenced, it was readily apparent that these scores should not be interpreted as the sole indication of an individual's leadership ability or capability. This follows the natural logic of the Infantry School's "whole man" approach to student evaluation and immediately concedes that there are multiple factors which should be considered when evaluating the leadership ability or potential of newly commissioned officers. Several factors, particularly those which consider leadership as essentially a relationship involving power and influence, have been evaluated by Fiedler in his contingency model interpretation of the effects of leadership training and experience. Fiedler refers to this interaction of power and influence as situation favorableness.[1]

The Infantry School is constantly exploring new dimensions of student evaluation with considerable attention to the area of situational leadership assessment. Recent efforts have resulted in the development of a combined peer (student associate) and preceptor (tactical officer) rating which may be of value as an additional dimension of a leadership development and evaluation plan of basic course officers. Within the overall scope of the Peer/Preceptor Rating Program, several combinations of factors in military leadership can be evaluated which could contribute uniquely to the assessment of junior officer leadership potential and perhaps assist in early leadership development and career orientation.

The purpose of this paper is to present some theoretical considerations, parameters, and pilot data generated during a trial run of the program.

THE RATING INSTRUMENTS

The program is based on two rating instruments; a peer (student associate) rating and a preceptor (tactical officer) rating.

"The Peer Rating"

The peer rating is a "nominative" type rating developed by the Army Research Institute.[4]

The rating is administered in platoon sized groups which normally range from 30 to 35 lieutenants; however, the design of the rating format will allow its use with groups as small as 13. Since the rating is a "nominative" rating, as opposed to a "total forced" rating, platoon members rate only their contemporaries whom they perceive as standing at the extremes in their leadership ability. Each student is required to identify those six individuals within his platoon whom he would "MOST" want in command of a platoon assigned to protect his flank in a combat situation and those six individuals within his platoon he would "LEAST" want assigned this same mission. All favorable, or "MOST", responses are equated to a numerical rating of 7, and all unfavorable, or "LEAST", responses are equated to a numerical rating of 1. Platoon members not receiving a "MOST" or "LEAST" rating from a specific individual are imputed a numerical rating of 4 as their rating from that particular student. Each platoon member must rate himself as a 4. Thus, each student will receive a rating of 1, 4 or 7 from every other member of his platoon. These ratings are averaged using the standard formula to derive the student's peer rating.

(SLIDE 1)

Since an individual platoon member must always rate himself a 4, it becomes obvious that his peer rating may range from a low of slightly more than 1.00 to a high of slightly less than 7.00 dependent upon the number of platoon members. In other words, extremely high or low ratings are asymptotic on 7.00 or 1.00. An average rating from all raters is 4.00.

In order to outline the parameters of peer ratings within a given platoon, the following illustrations are presented. (These illustrations are based upon a platoon population of 33 members.)

(SLIDE 2)

1. Students are in agreement on those who should receive all the "LEAST" (1) ratings, and those who should receive all the "MOST" (7) ratings within the platoon. This agreement will fix the top and bottom ends of the platoon distribution; six individuals will receive a maximum score and six individuals will receive a minimum score. (This maximum and minimum score will vary with the size of the platoon.) The center of the distribution will also be fixed as the remaining 21 students will receive a rating of 4 from every other student and, therefore, attain an average rating of 4.00.

(SLIDE 3)

2. Students are in agreement on those who should receive all the "MOST" (7) ratings, but cannot agree on those who should receive all the "LEAST" (1) or average (4) ratings.

The agreement of the students will fix the top end of the platoon; six individuals will receive a maximum score. However, since the members cannot agree on those individuals who should receive 1 or 4, the remainder of the distribution is left to fluctuate. (This distribution is reversed when the students can agree only on those six students who should receive all the "LEAST" (1) ratings.)

(SLIDE 4)

3. If the students are not in agreement with any of their ratings, the mean of the ratings received by each student will be 4.00. The distribution then becomes a point. When the data conform to this distribution, it may be suspected that the students have collaborated and developed a cycle to "beat" the system.

"THE PRECEPTOR RATING"

The preceptor rating is an assessment of each individual platoon member's leadership qualities and performance by the platoon tactical officer (an experienced officer usually in the rank of captain). This assessment uses a standardized subjective rating format and is based on 8 so-called performance factors and 12 leadership qualities. Within the program, this rating is entitled the "Performance Evaluation Report." These performance factors and leadership qualities were evolved from a rating scale previously utilized in the Infantry Officer Candidate Program. Through the expertise and consensus of several experienced tactical officers and cadre company commanders, this original format was modified and adopted to the IOBC Program.

(SLIDE 5)

The student receives a numerical score ranging from 1 to 5 in each of the rating areas.

A student's preceptor score is computed as the sum of each numerical score received in all of the performance factors and leadership qualities. This score will range from a low of 20 to a high of 100, a purposeful skew to facilitate score interpretation and allow these scores to more closely approximate a traditional grading scheme.

THE PROGRAM

The leadership rating and evaluation program is specifically designed to allow officer students and tactical officers to participate jointly in the leadership evaluation process.

Based upon a projected input of 200 students per class, each Infantry Officer Basic Course is organized as a student company with six platoons of approximately 33 students each. Leadership development begins upon initial inprocessing with the assignment of a student chain of command and continues throughout the course. Positions within the chain of command for each company are as follows.

(SLIDE 6)

By rotating positions weekly, a total of 468 leadership positions are generated during the 12-week course. The standards for leadership development, to include the performance objectives of each leadership position, are outlined in the IOBC Officers' Guide which is issued to each officer at the beginning of the course.

Leadership evaluations, both peer and preceptor ratings, are rendered twice during the course. As leadership positions are rotated weekly, the first rating does not occur until the 7th week.

(SLIDE 7)

By the 7th week, each student has been placed in a leadership position exposing him to both his peers and the tactical officer. This evaluation consists of both a peer and preceptor rating combined at a ratio of 2/3 peer to 1/3 preceptor. A second evaluation is rendered at the 11th week. This 11th week evaluation is computed in the same manner as the 7th week evaluation: 2/3 peer to 1/3 preceptor. After the 11th week evaluation has been made, the student's final leadership score is computed by combining the 7th and 11th week ratings at the ratio of 1/5 and 4/5 respectively.

(SLIDE 8)

The student's final leadership score is then implemented into the total IOBC student evaluation process by combining it with the student's final academic grade at the ratio of 1/3 leadership to 2/3 academic. This determines the final order of merit listing for the course.

Since an individual's peer rating could vary from slightly above 1.00 to slightly less than 7.00, a method was developed to convert these scores to a scale compatible with the preceptor scores. This method also established the peer ratings on a more traditional grading scheme with a converted score of 70 becoming the minimum passing score. This conversion, which utilizes a norm-referenced concept, is a simple process and involves the same basic principles as for converting yards to meters and vice versa. Fundamentally, it consists of locating the equivalent points on the two scales and establishing an arithmetic ratio between the two ranges of scores. For purposes of the pilot study, the reference points on the peer scale were located _a priori_ at 2 standard deviations below and 2 standard deviations above the mean (4.00) where n = 500. (When n becomes very large, the distribution approaches normality even though the ends are fixed.) These reference points are then equated as minimum passing, 70, and minimum superior, 95, scores on the converted scale. This concept is best illustrated by the conversion graph.

(SLIDE 9)

TEST DATA

A preliminary test of this program was conducted at the U. S. Army Infantry School in September 1973. However, the program was not implemented in strict accordance with its original design. Personnel resources within the Infantry School were critical and only 3 officers possessing the necessary qualifications could be made available to act as tactical officers.

Despite this limitation, the pilot study provided an excellent tool for management by exception as it alerted the tactical officers, through the peer rating, to those students in possible need of leadership counseling or additional training. Since the pass/fail standard of the peer rating for each platoon is determined through a norm-referenced procedure, those students who fall below the criteria established within the platoon can be very easily identified. Most of the distributions of the pilot data are considerably flatter than the parameters previously outlined; however, the basic distributions of the pilot data do conform to these parameters.

(SLIDE 10)

The center of the pilot distribution approaches that of the parameter. The extremes are relatively stable but not as "fixed" as those of the illustration.

(SLIDE 11)

There appeared to be some agreement among the students as to who should receive predominately "MOST" (7) ratings. The remainder of the distribution tapers off toward 1.00.

(SLIDE 12)

This distribution seems to point out a disagreement among the students as to who should receive all "MOST" (7) or all "LEAST" (1) ratings. Note the "shrinkage" of the distribution as the extreme scores have closed in toward 4.00.

It is expected that the greatest contribution of the program to student evaluation is its identification of those students who had below standard leadership scores but whose academic scores were above the standard required for graduation. The Leadership Assessment Program appears to have revealed an entirely new dimension to student evaluation-- a tool to identify those students requiring additional training or direction in their leadership style.

Since the pilot program was conducted with less than a full complement of tactical officers, it would be inappropriate to draw any conclusions from the data at this time. It was not the intent of this paper to offer any conclusive evidence of how such instruments should or should not be used, but simply to present the combined instrument as a possible indicator for leadership development and evaluation.

SOURCES

1. Fiedler, Fred E., The Effects of Leadership Training and Experience: A Contingency Model Interpretation. *Administrative Science Quarterly*, 1972, *17*, 453-470.

2. Haggerty, H. R., *Status Report on Research for the U. S. Military Academy*. Technical Research Report 1133. U. S. Army Research Institute for the Behavioral and Social Sciences, Arlington, Virginia, October 1963.

3. Tobin, D. J., Major and Macrum R. H., Major, *Leadership Evaluation*. Research Report, Office of Military Psychology and Leadership. West Point, New York, 1967.

4. USARIBSS Research Note, later published as: Downey, R. G., *Associate Evaluations: Nominations vs Ratings*, Technical Paper 253. U. S. Army Research Institute for the Behavioral and Social Sciences, Arlington, Virginia, September 1974.

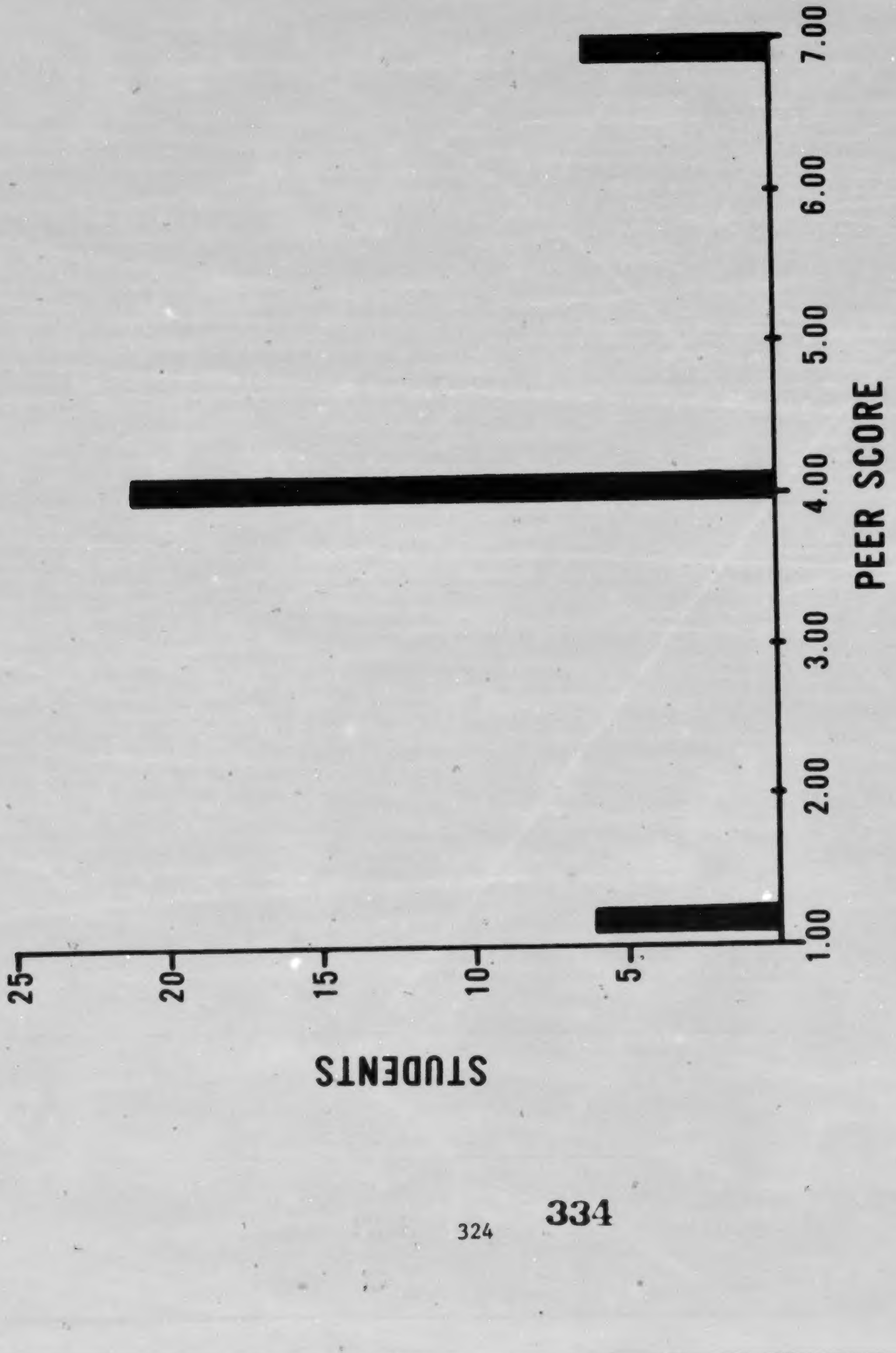
COMPUTATION OF PEER SCORES

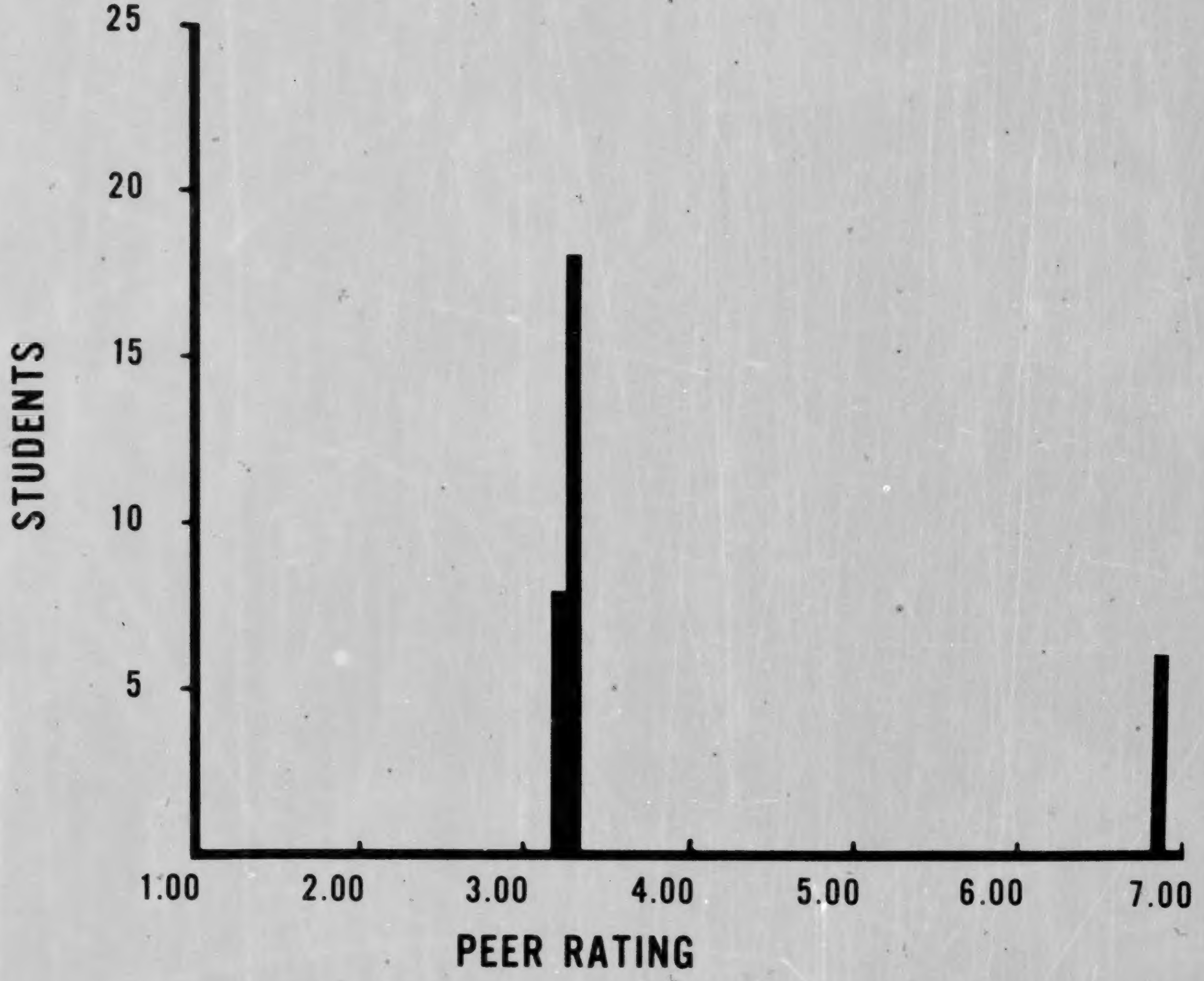
$$R = \frac{\sum_{1}^{n} r}{n}$$

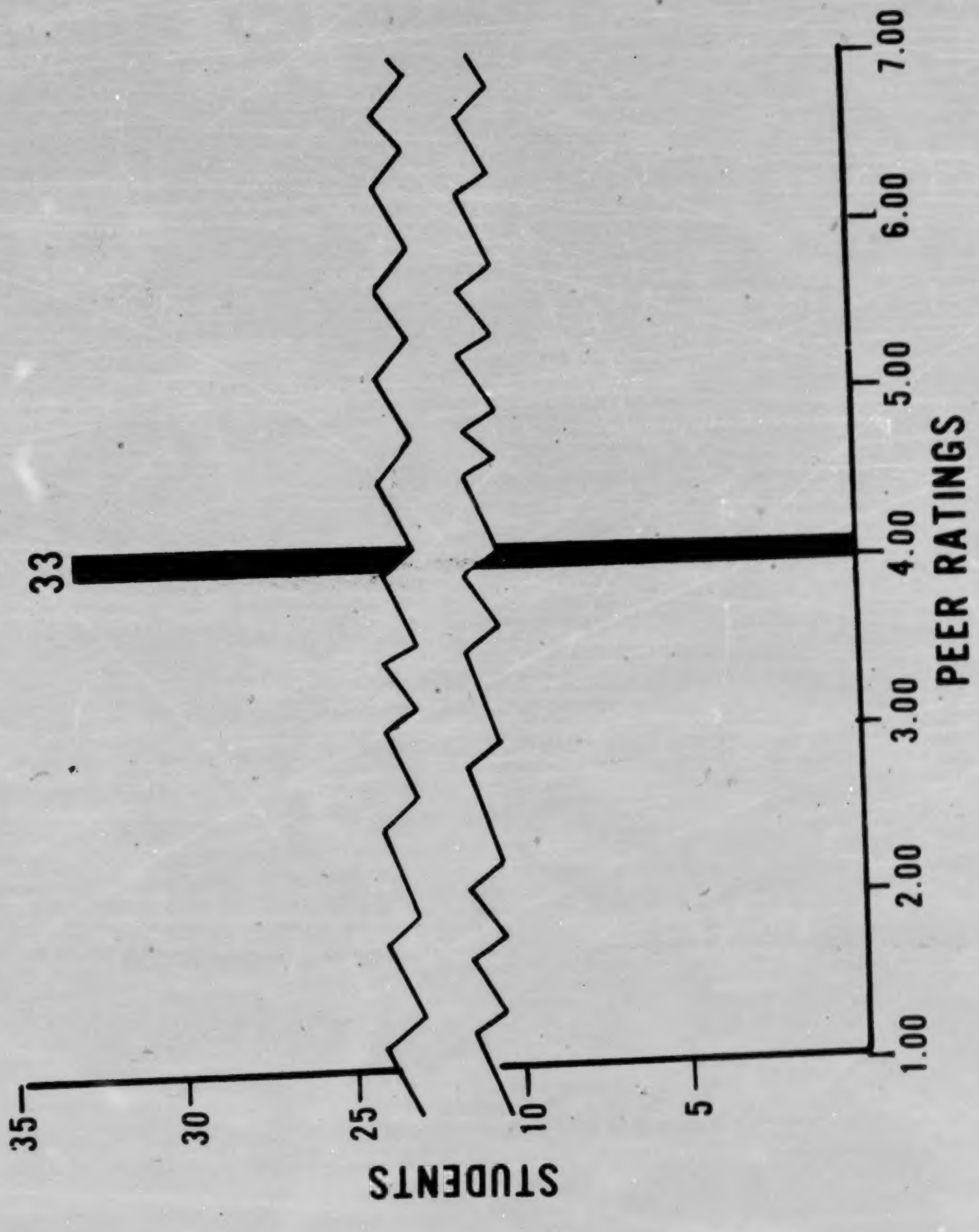
WHERE: R = PEER RATING

n = NUMBER OF INDIVIDUALS IN THE PLATOON

r = PEER RATING GIVEN BY A SPECIFIC PLATOON MEMBER


STUDENTS
25
20
15
10
5
PEER SCORE
1.00
2.00
3.00
4.00
5.00
6.00
7.00


25
20
15
10
5
STUDENTS
1.00
2.00
3.00
4.00
5.00
6.00
7.00
PEER RATING


STUDENTS
35
30
25
10
5
33
1.00
2.00
3.00
4.00
5.00
6.00
7.00
PEER RATINGS

PERFORMANCE FACTORS

PROFESSIONAL KNOWLEDGE
QUALITY OF PERFORMANCE
DELEGATES AUTHORITY
SUPERVISION
RESPONSIBLE FOR ALL ACTIONS
WELFARE OF SUBORDINATES
EXPRESSION
PLANS FOR FUTURE OPERATIONS

LEADERSHIP QUALITIES

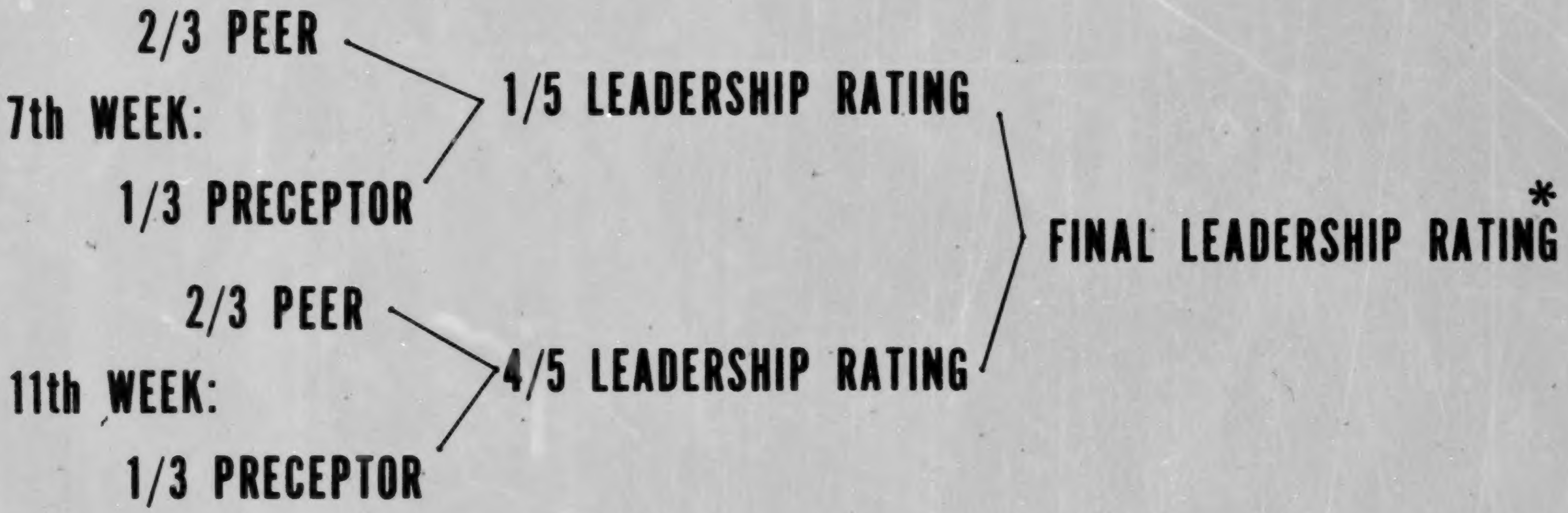
ADAPTABILITY
BEARING
COOPERATION
DECISIVENESS
DEPENDABILITY
ENTHUSIASM
FORCE
INITIATIVE
JUDGEMENT
INTELLIGENCE
SELF-IMPROVEMENT
TENACITY

IOBC

STUDENT CHAIN OF COMMAND

POSITION	NUMBER PER COMPANY
COMPANY COMMANDER	1
EXECUTIVE OFFICER	1
FIRST SERGEANT	1
PLATOON LEADER	6
PLATOON SERGEANT	6
SQUAD LEADER	24
TOTAL	39

LEADERSHIP SCORES



7th WEEK:
- 2/3 PEER
- 1/3 PRECEPTOR

→ 1/5 LEADERSHIP RATING

11th WEEK:
- 2/3 PEER
- 1/3 PRECEPTOR

→ 4/5 LEADERSHIP RATING

1/5 LEADERSHIP RATING + 4/5 LEADERSHIP RATING → FINAL LEADERSHIP RATING*

* STUDENTS WITH A FINAL LEADERSHIP RATING BELOW 70% WOULD BE SUBJECT TO FACULTY BOARD ACTION

FINAL CLASS STANDING

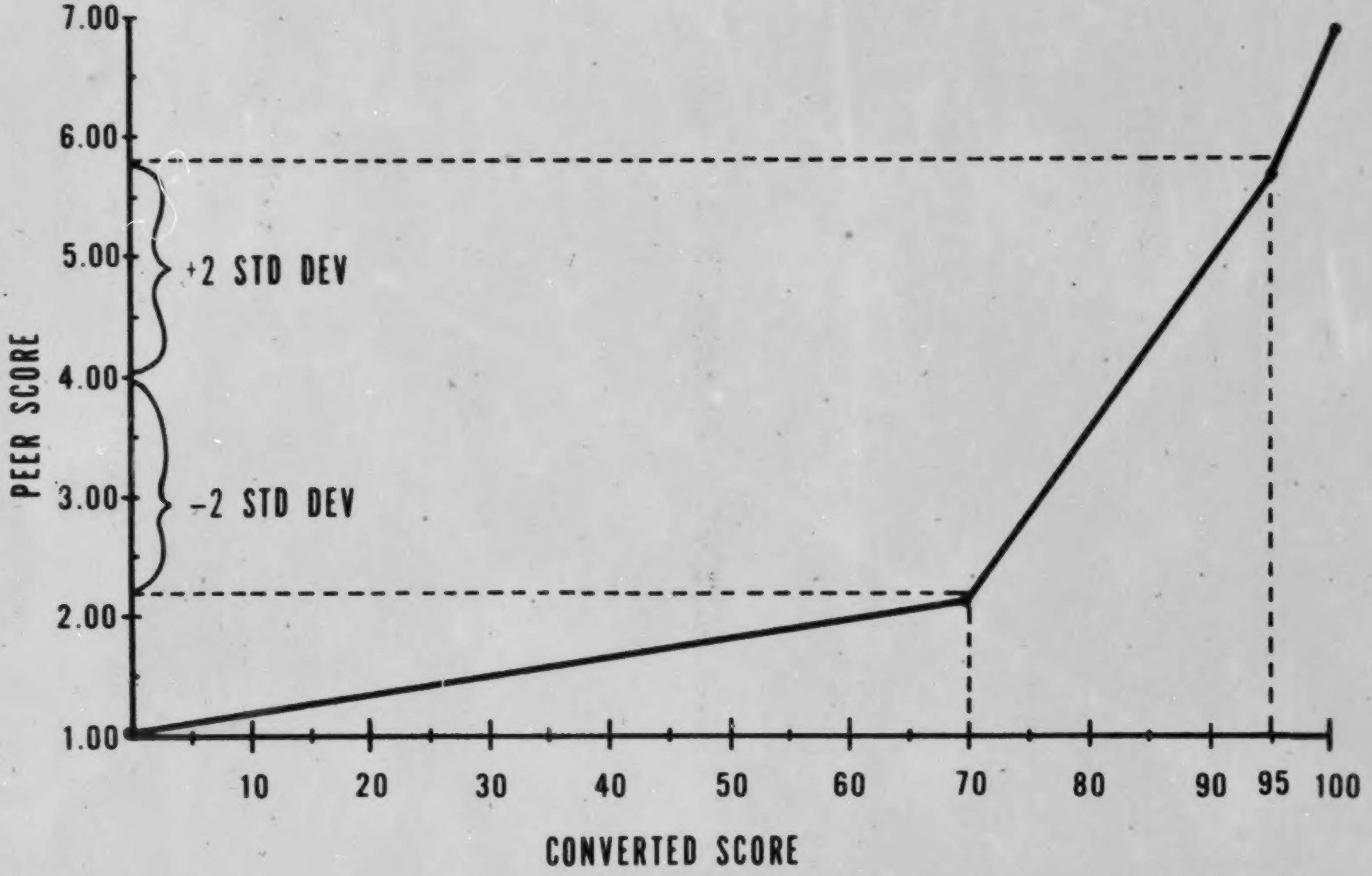
1/3 FINAL LEADERSHIP RATING

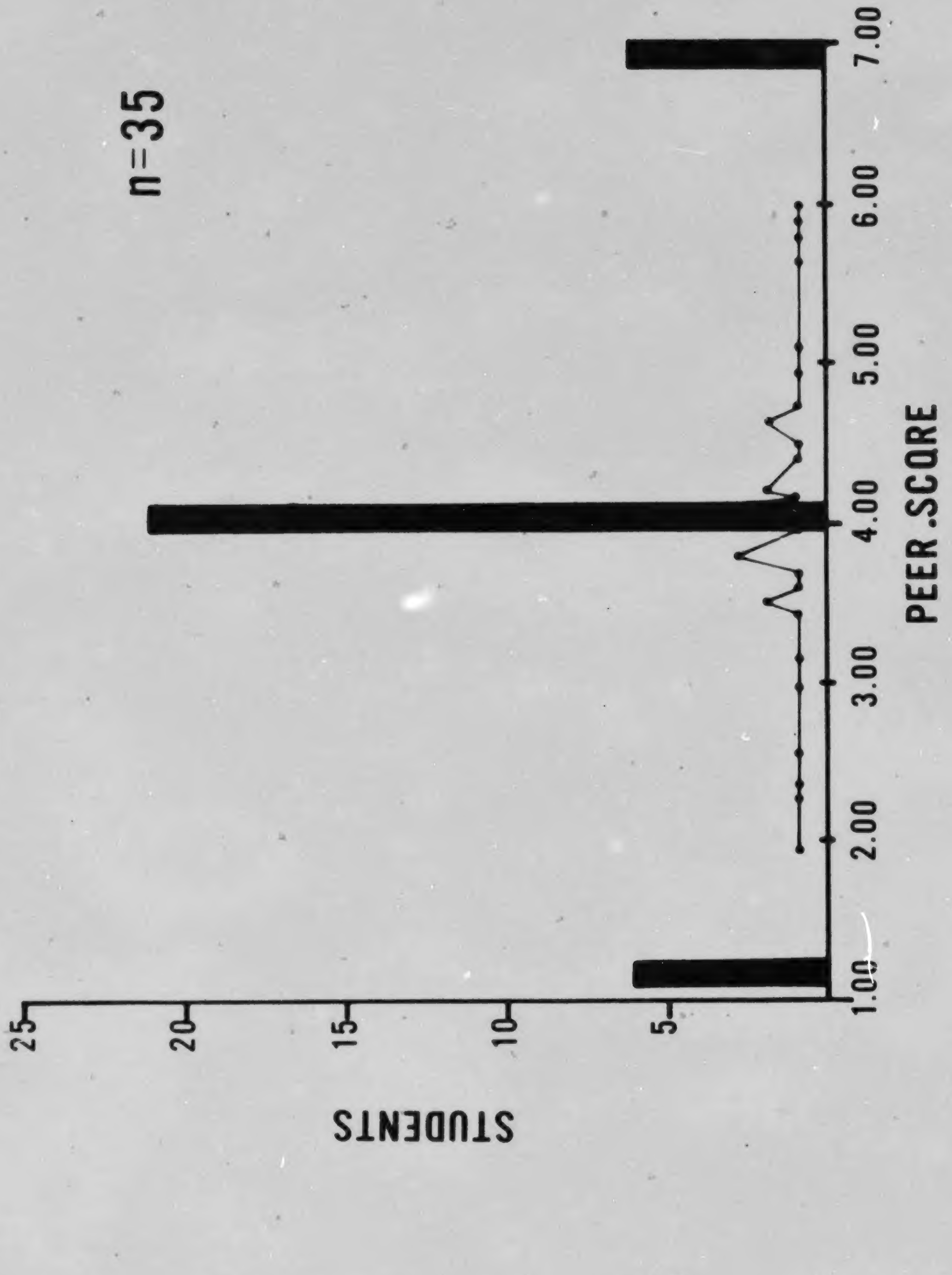
FINAL CLASS STANDING

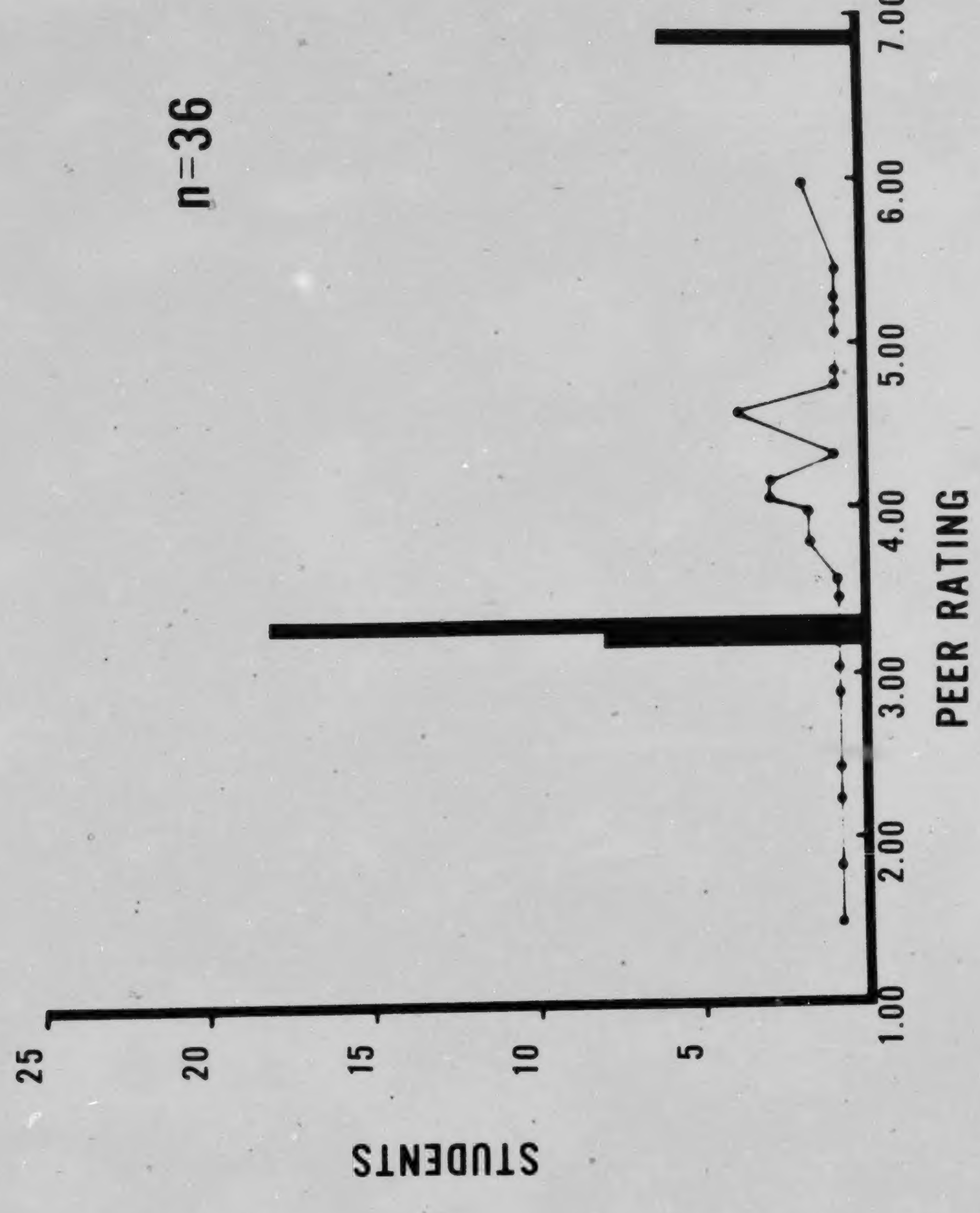
2/3 FINAL ACADEMIC AVERAGE *

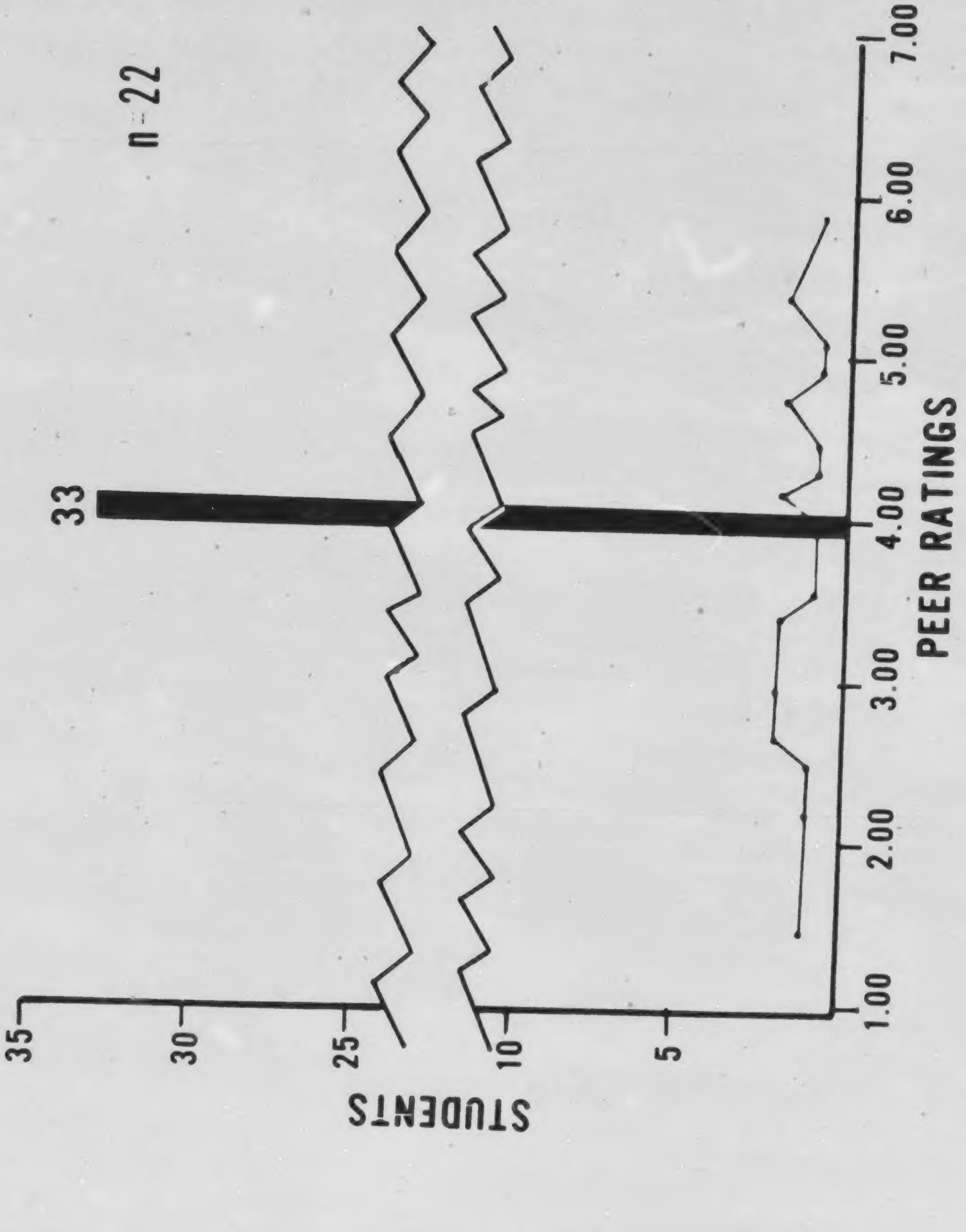
*STUDENTS WITH A FINAL ACADEMIC AVERAGE BELOW 75% ARE SUBJECT TO FACULTY BOARD ACTION

PEER SCORE CONVERSION GRAPH




n=35
STUDENTS
25
20
15
10
5
1.00
2.00
3.00
4.00
5.00
6.00
7.00
PEER SCORE


n=36
STUDENTS
25
20
15
10
5
PEER RATING
1.00
2.00
3.00
4.00
5.00
6.00
7.00



n=22
33
STUDENTS
35
30
25
10
5
1.00
2.00
3.00
4.00
5.00
6.00
7.00
PEER RATINGS

VOCATIONAL ADVANCEMENT IN THE FEDERAL ARMED FORCES

LT COL J. FRICKE

Gentlemen,

Thank you for the invitation and for giving me the privilege to speak before this meeting of the Military Testing Association.

I am not here as the representative of our Ministry of Defense but as the chairman of the German Air Force on the board of chairmen - the presidium - of the Federal Armed Forces Association (the Deutscher Bundeswehr-Verband).

Ours is an association of active soldiers with about 180.000 members. The Association is not a union.

As partner of the military authorities we represent our servicemen's ideal and social interests in the political field. Financially and politically we are absolutely independent.

Amongst our members are Privates - 6.000 of them - as well as NCO's, officers and highest ranking admirals and generals of all service branches, including the Chief of Staff of the Federal Armed Forces, Admiral Zimmermann.

Gentlemen, you may find some words or terms in my presentation a little hard to understand. There are apparently a number of things unique to our system, respectively no generally agreed translation exists. Please feel free to ask if such a term occurs.

1. Who is eligible for Vocational Advancement?

The Federal Armed Forces have a total strength of 465.000 servicemen. Of these, 230.000 are draftees (conscripted soldiers), 180.000 long term volunteers, and 55.000 professional soldiers (regulars). The 180.000 long term personnel fall under the provisions of the Vocational Advancement Program.

Annually, about 20.000 long term volunteers with enlistment terms of 4 to 15 years are discharged from service. Since inception of the Federal Armed Forces in 1956, approximately 200.000 long term volunteers have received vocational advancement.

2. What does the Vocational Advancement Program offer?

a) Financial benefits

Financially, long term personnel upon discharge receive a single mustering-out payment or transition allowance, consisting of between 3 times and 14 times the last monthly pay, and recurrent payments up to three years after discharge (transition pay), amounting monthly to three quarters of the last monthly pay drawn in service. Amounts and duration of payments depend upon length of service. In addition, 60 per cent to expenses are taken over (reimbursed) in event of sickness.

Two examples on financial benefits, based on figures for 1974: A Sergeant First Class, married, two children, receives after 12 years of service a total of around DM 75.000 (U.S.: $30.000). A Captain (Army rank), also married and with two children, would be paid over DM 100.000 (U.S.: $40.000) for his 12 years with the Armed Forces.

b) Vocational Advancement Training

(a) General information

Besides the financial benefits, the long term volunteer is given vocational advancement training of between 6 months and 4 1/2 years - depending on the length of his enlistment.

For enlistments of 8 years, the Vocational Advancement Program begins during the last year of service - the serviceman being relieved from active duty - and for those of 12 years already during the last 1 1/2 years. Active service in these cases is thus shortened to 7 respectively 10 1/2 years.

For vocational advancement, the soldier can choose between going to one of the Armed Forces Vocational Schools for general education and a special vocational training at normal civilian training facilities.

The Vocational Advancement Service carries the expense of this program, which means an additional indirect financial benefit of DM 9.000 (U.S.: $3.600) for the time covered by the program.

I would like to point out that since a few years long term officers are also eligible under this program, as well as jet pilots retired at the age of 40 - although with slight modifications. Recently, too, parts of the system have been opened to our draftees. Since figures on these categories are not yet available I just mention the fact.

(b) Reform of Education and Training

In context with a recent reform of the Education and Training Program of the Armed Forces, parts of the Vocational Advancement Program also have been remodelled. Was vocational advancement training formerly confined to the end of the time in service and to the time after discharge, it now also takes place during the military training period - civilian vocational training with civil authority acknowledged graduation being given at the same time and as a part of it. This means, for instance, that a soldier with a relatively short term of enlistment (up to 8 years) during his training for Senior NCO (Staff Sergeant) also is taught to become a master craftsman. Correspondingly, a man with a long term enlistment of 12 years during his military training can gain the level of graduated technician, industrial engineer and social worker.

(c) Vocational Guidance

Specialists of the Vocational Guidance Division, part of the Vocational Advancement Service, advise the servicemen about possibilities in the various professions prior to the begin of the advancement program. On the average, every eligible soldier is contacted and advised by this division every 2 years. The Vocational Guidance Division - as in fact the whole organization - closely cooperates with the units, the Armed Forces Vocational Schools, and the Federal Employment Agency.

(d) Organizational scheme of the Vocational Advancement Program

The Ministry of Defense has under its authority the
- Armed Forces themselves
  and the
- Armed Forces Administration.

The Vocational Advancement Program is under the jurisdiction of the Armed Forces Administration and is made up of two organizations or services.
- The Vocational Advancement Service, responsible for vocational guidance and training of servicemen. There are 36 branch offices of this organization at the various subregion recruiting offices.
- The Federal Armed Forces Vocational Schools, offering general educational advancement up to (German) High School Graduation Diploma (Abitur: Level up to two years of college).

(e) Facts and Figures

I now would like to present to you some statistics on scope and results of vocational advancement programs carried out so far by the Armed Forces.
- In programs conducted on a voluntary basis during off-duty hours or during the serviceman's active duty time, since 1960 more than 15.000 long term personnel were trained to the level of journeyman, assistant and clerk; over 3.000 passed the examination for master craftsman during their service career.
- Since 1960, vocational training was approved and conducted for the following numbers of servicemen:

Training	in 1971	in 1972	since 1960
to master craftsman or mechanic	1.555	1.635	14.780
to technician	2.301	1.875	13.543
to engineer	712	596	4.424
in welding	451	287	5.734
in mechanical engineering	128	95	2.525
in commercial professions (to medium-level management)	1.928	1.275	12.348
in foreign languages (to translator, foreign language clerk, etc.)	233	183	2.252
in the data processing field	444	311	5.160
in the medical and health services field	580	504	3.211
to teacher	453	508	2.238

- More than 100.000 servicemen have gone through Armed Forces Vocational Schools and completed education courses.

These figures may show you the importance and value of the Armed Forces Vocational Advancement Program in the field of adult education.

3. Efforts and achievements of the Federal Armed Forces Association

It would lead too far to list here all efforts - and also achievements - of the Federal Armed Forces Association in this field. But vocational advancement and the successful civilian reorientation of the numerous long term personnel annually up for discharge - of which no less than 80 per cent are members of the association - is one of the main objects of the association's work and one of the main problems, too.

The Association has backed up its demand for a frictionless and unproblematic transition or reentry of discharged long term personnel into civilian professions with numerous studies on the subject, extensive questionnaire programs, law-suits up to the highest instance in Administrative Court (Verwaltungsgericht), and of course through its efforts and possibilities in the political field towards legislature. Substantial improvements in respect to pay, social care and other benefits and especially in the field of vocational advancement are the fruit of permanent efforts by our association.

In closing, permit me to point out that the Vocational Advancement Program of the Federal Armed Forces may be called one of the best, at least as far as Europe is concerned. This program of our Armed Forces has also served as model for the civil laws on vocational education in the Federal Republic, enacted 1969/70 in regards to adult education, such as the Employment Promotion Act (Arbeitsfoerderungsgesetz) and the Vocational Training Promotion Act (Ausbildungsfoerderungsgesetz); a multitude of vocational improvement programs and instruction in schools and universities also benefitted from procedures and experiences of the Armed Forces.

SECTION IX

SYMPOSIUM:

OCCUPATIONAL ANALYSIS IN THE MILITARY ESTABLISHMENTS

OCCUPATIONAL ANALYSIS IN THE UNITED STATES AIR FORCE

by

Walter E. Driskill, Ph.D.
Chief, Occupational Survey Branch
United States Air Force Occupational Measurement Center
Air Training Command
Lackland AFB, TX

Opening Remarks

In this symposium we are addressing the current state of the art in occupational analysis in the armed forces of this and other nations. To provide a frame of reference I wish to point out that our programs have some essential similarities: A task list or job inventory approach to data collection, large samples, computer analysis of raw data, and application to a wide variety of management systems, two of the foremost being personnel classification and training. In the case of the US Armed Forces, we are all using (or will shortly use) the Comprehensive Occupational Data Analysis Programs, usually referred to as CODAP, developed by the Air Force.

There are, however, some fundamental differences, having to do with program purposes, data collection techniques, and philosophy of analysis. The intent of this symposium is to compare and contrast our separate programs.

Air Force

In this presentation I describe the Air Force's operational occupational analysis program that is conducted by the Air Training Command for the United States Air Force. Occupational analysis provides us with information that is used both for descriptive purposes and for decision making purposes. Occupational analysis as we define it in the Air Force consists of the detailed specification or listing of the tasks that may be performed by airmen in a particular occupational field and the determination of the percentages of airmen who perform each of these tasks. Through the procedure we also collect large amounts of biographical information about job incumbents and information about their work environment, such as the kinds of aircraft they work on, the kinds of equipment they maintain, and their satisfaction with their jobs. A third kind of information that is useful to us is information about how Air Force occupational fields are actually structured. We define in Air Force Manual 39-1, jobs in each of the occupational fields, but these descriptions are based on what we believe is done in the fields and they are organized according to how we believe they are structured. Occupational analysis, through our complex grouping analysis program, provides us with information on how jobs are actually being performed and how they are organized.

Research into the Air Force occupational analysis program began in 1956 and now 17 years later it has grown into a large and very useful operational program. By 1965 the methodology, including the computer programs for analysis, were ready for operational implementation but it was not until July of 1967 that the operational program began in Air Training Command. At that time the program was geared to surveying 12 occupational fields annually and consisted of 15 professional, technical, and clerical personnel. The following year the program was expanded to a capability of surveying 24 occupational fields annually with 28 personnel. Just last year the program was again expanded to where we are now capable of surveying 51 fields annually with a total manning of 41. Since the beginning of the program in 1967, surveys of 125 fields have been completed and 60 more are under study.

The occupational analysis process of the Air Force consists of four distinct steps. The first of these is the development of the job inventory or description of tasks that may be performed in an occupational field. The second step consists of the validation of this list of tasks with subject matter specialists in operational units worldwide. Once the task list is validated, it is then administered to job incumbents through Consolidated Base Personnel Offices worldwide. The final step really consists of two parts, the first consisting of the computer analysis of the data and the second of the manual analysis of what the information means in terms of Air Force classification, assignment, and training. In the development of a job inventory, our inventory development specialists research all of the reference material pertinent to the occupational field being surveyed. From this research they develop a tentative list of tasks which they use in face-to-face interviews with subject matter specialists, first at the technical training centers with the responsibility for training in the specialty or occupational field, and second in operational units where subject matter specialists are actually doing the jobs. Once a point of diminishing returns in interviews is reached, the task list is reproduced and mailed to subject matter specialists in operational units in the continental United States as well as overseas for validation.

The comments and changes from the field validation of the task list are incorporated in the final task list which is then published as a United States Air Force Job Inventory. Here, I should like to say that occupational analysis is largely a matter of communication. A case in point is that we have been operating for over six years using the United States Air Force Job Inventory as our job instrument, and the Personnel Research Division of the Human Resources Laboratory used this particular instrument for a number of years prior to our operational implementation. It was just recently, however, that we found that the term job inventory does not communicate with job incumbents in many of our occupational fields, so now we are in the process of searching for another title for our task list and biographical information section.

The job inventory consists of two divisions. The first section consists of questions to elicit background and work information about job incumbents. We obtain from job incumbents such information as their name, grade, social security number, occupational field, and how long they have been in their career field, and their present job. In addition, we obtain work environment information, such as the kinds of equipment they work on, their satisfaction with their job, and the kinds of aircraft that they may work on. The amount of information that can be obtained through the background information section is limitless and is extremely useful for making decisions about the jobs that personnel hold.

The second section of the job inventory consists of a detailed listing of tasks that airmen may perform. In completing this inventory, job incumbents read through the tasks that are listed, checking the tasks that they do in their present job. Once they have completed this procedure they go back and rate such task according to the amount of time they spend on it relative to the others they perform.

The task is administered by local Consolidated Base Personnel Offices to personnel specified by us. We use the Uniform Airmen Record file provided by the Military Personnel Center and specify to the Base Personnel Offices the numbers of personnel of various kinds to whom they are to administer the inventory. Our sample is a random stratified sample by skill level, command, and job location, and the total number depends upon the size of the career field. In those with 3,000 or less incumbents, we obtain a total sample, that is, we try to administer the inventory to every individual who has been on the job over six weeks. For occupational fields with more than 3,000 incumbents, some percentage of that total number is obtained. The largest survey that we have accomplished, for example, is in the Aircraft Mechanics, occupational field in which over 70,000 job incumbents were working at the time of the survey. We obtained a sample of about 12 percent of those personnel.

The job inventories are returned to our organization where we carefully scan and review the returns and prepare the data for computer processing. The data are then input into the Human Resources Laboratory computer where they are analyzed through the Comprehensive Occupational Data Analysis Programs, commonly called CODAP. CODAP is nothing more than a series of highly complex computer programs that are used to reduce the large amount of data obtained from the job incumbents in the field into a manageable form.

As a result of the CODAP analysis we obtain two kinds of information. The first consists of biographical and work information about the job incumbents or various categories of job incumbents. We are finding useful such information as location of assignment, education level, equipment utilization, reenlistment potential, and job satisfaction. In other cases, depending upon the occupational field being surveyed, other information is extremely valuable.

The second kind of information pertains to the tasks and percentages of airmen performing these tasks. While CODAP has variety of ways of manipulating the data, the important thing to remember is that whatever the manipulation there are, these kinds of basic information provided: first, a listing of each of the tasks performed by airmen in the occupational field; second, the percentages of any given group of airmen performing each of the tasks; third, the percentage of time spent by the members who perform the tasks; fourth, the percentage of time spent by all members of the group averaged across all members of the group; and fifth, an accumulation of the time spent on the tasks.

The Air Force has put the data to use in a variety of ways, among the more important are these: career field structuring, in which the organization of occupational fields are validated or restructured; the determination of training requirements and instructional system development, which I will talk about a little later: for personnel research where our data are being used consistently by Dr. Christal and his staff who answer research problems; or the determination of the kinds of maintenance engineering policies that should exist in occupational fields; and for the development of tests used to promote airmen in grades E-4 through E-7.

One of the major uses of the data is in the instructional system development model usually called ISD. This model consists of five steps: analyze systems requirements; define education or training requirements; develop objectives and tests; plan, develop, and validate instruction; and conduct evaluation of instruction. Occupational data are used in the first step to analyze system requirements as a basis for defining education and training requirements as well as in the last step in the conducting of evaluation of instruction. We have found, for example, that where we have already had an instructional system development project in an occupational field prior to an occupational survey, the survey results are very useful for evaluating the efficiency and effectiveness of the course of instruction that resulted from the ISD application.

A most important use of the data is in determining training requirements. I would be remiss, however, if I were to lead you to believe that we immediately made significant advances in determining training requirements upon the initial use of occupational data. Rather, it was after long and extensive use of the data that we came to know how to make the fullest use of results. While to the vast majority of the people who first use it, occupational data appears to be very useful. When the time comes for them to make use of the data they simply are unable to make decisions without the use of some criteria that is provided by a policy making agency. Here, I should emphasize that I feel that any organization that uses occupational data is going to have to define some decision criteria and that these decision criteria are fundamental to the effective use of the data.

To show you how we came upon the decision criteria, let me illustrate some of the kinds of things that we found in the survey results. Here is a specific example in a highly complex occupational field dealing with the maintenance, repair, and calibration of precision measuring equipment, such as oscilloscopes and scintillation counters. We found that there were some tasks that most, or a majority, of the job incumbents performed. In other words, for each 100 percent of the people that we trained we found use of the training by 60 percent or more when the course graduates went to their first job. We found, on the other hand, many tasks performed by small percentages of the job incumbents. In other words, for every 10 people that we trained, we had three or less using their training in their first job. This situation occurs in all career fields in the Air Force, and, in fact, there are some occupational fields in which no tasks are performed by very many personnel, while in others there are many tasks performed by many people.

This led us in Training Command to reaffirm a concept for training which emphasizes providing initial training on the tasks for which the probability of performance by airmen in their first job assignment is high. On tasks for which the probability of performance is low, other means of training, through Career Development Courses, through on-the-job training, or through advanced or supplementary training, should be used to develop proficiency.

Even this policy statement was insufficient to get occupational data used for decision purposes, and it was necessary to levy some specific criteria for determining training requirements. Specifically, when an occupational survey of a field is completed, Training Command provides the training center with responsibility for training with these criteria: for tasks performed by 50 percent or more of the job incumbents in their first job, training is to be provided so that very minimum OJT is required for the course graduate on his first job. For tasks performed by between 30 and 49 percent of the job incumbents in their first assignment, background or fundamental training is to be provided. The intent of this training is to make job proficiency training by supervisors easier. For tasks performed by less than 30 percent of the personnel in their first jobs, training is to be reduced or eliminated, unless it can be shown that it is cost effective to train everyone in order to train the few who will use their training in their first job.

The application of these guidelines has resulted in very significant training cost avoidance. Here is an example. In the precision measuring equipment specialty at Lowry Technical Training Center, the course was 43 weeks in length. As a direct result of the application of occupational data, the course was reduced to 32 weeks, representing a 25 percent reduction in course length and a dollar "savings" of over one million in one fiscal year. Other significant course reductions have been made in other specialties. Overall, as a direct application of occupational analysis data to determining training requirements and as a part of instructional system development program, over 12 million dollars can be shown to have been saved in the past six years.

I should like to conclude by pointing out that not all surveys have resulted in the significant savings that I have just shown. I can say categorically, however, that every occupational survey has resulted in training benefits. In the first case, and this fact is frequently overlooked by training management, we have been able to validate training. It is much better to find that the training you are providing is on target than it is to find that it is off target. In the second case, we have frequently been able to reallocate training time within an existing course. While the total course length has not been reduced, we have been able to give more emphasis to some areas deserving it and to give less emphasis to other areas. Finally, as you have just seen, we have been able to reduce training from a very small amount, a few days, to very large amounts, consisting of weeks of training. I am confident that the use of occupational survey data in any training program can result in a more effective and efficient training program regardless of the environment. Thank you for your time.

NAVAL OCCUPATIONAL TASK ANALYSIS PROGRAM

By

M. D. Callahan

Two years ago at the annual MTA meeting I gave a two-hour presentation on the Naval Occupational Task Analysis Program known as NOTAP. I was programmed for only one hour but a gentleman from overseas, for whom I substituted, never showed. I note that our foreign colleagues in occupational analysis made it this year, including our old friend from Canada, CDR Cormack.

During the presentation I made the statement that NOTAP, although still funded under R&D, was funded for the operational phase, commencing 1 July 1973. Upon returning to Washington, I found we were very much unfunded. However, we have been operational since 1 July 1973 and have a vigorous, ongoing program with strong support from the Bureau of Naval Personnel--this time, though no comments about funding!!

I will spend about three minutes reviewing the mission, operation, organization and use of NOTAP data and then, as Dr. Driskill requested, spend the remainder of the time in making a few interservice comparisons. (See Figure 1.)

First, our mission--this mission has been reinforced by a recent Bureau of Naval Personnel Notice, distributed Navy-wide. (See Figure 2.)

Next, our organization--not the four divisions and that we report to the Chief of Naval Personnel for program guidance. (See Figure 3.)

How do we operate--after being assigned a Navy rating to survey and analyze, we would in general, routinely follow the steps under each phase as indicated on this slide. (See Figure 4.)

And last, the payoff--Who uses our data? (See Figure 5.)

Recently, a high level study group completed an in-depth analysis of the enlisted classification (rating) structure. While still in the implementing stage, certain significant changes in concepts and career structure were recommended. Where available, NOTAP occupational data was used in the planned implementation of the study group recommendations.

With regard to the NEC's, I will cover them later.

Training has been our largest user to date and they routinely receive certain printouts on each rating as they become available. We are grateful for strong support given NOTAP by the Training Command.

With reference to the qualifications for advancement, or occupational standards, the policy has been established that NOTAP data will be one of the primary inputs to that program.

It is well to remember:

1. Standards are what should be.

2. NOTAP is what is at a particular time.

3. The Chief of Naval Personnel determines these occupational standards.

4. The Chief of Naval Training trains to these standards and writes examinations based on these standards.

5. NOTAP data serves both; in the future, by inserting these standards as task items, along with other tasks, it will be possible to evaluate the currency and adequacy of these occupational standards.

An interesting observation on this two-way information flow loop--there are inherent checks and balances in the circuit.

PART II

SLIDE 1. (See Figure 6.)

And now for a few interservice comparisons--with some comments on why the Navy may or may not be doing certain items of the program, or may be doing them differently.

Such programs must be tailored to the needs and requirements of the particular organization, so what is necessary or good for one service may not be for a different service.

Likewise, the two slides we constructed are "busy" and it was necessary to make generalizations where, perhaps, some finer distinctions may be more appropriate. After CDR Nicholl's joke about the Irishman, I understand why I was tasked to make these interservice comparisons--I note that while you can see the next two slides, fortunately, the gentleman whose programs I am comparing, cannot see them from their present positions.

A. Prior to preparation of the inventory for a particular rating, the team concerned conducts an observation/interview with personnel on the job of a small, but representative sample of the rating (from 1% to 3%). We consider this valuable in developing a better inventory as it has pinpointed subsequent sampling or problem areas within a rating. Our inventories are constructed primarily by enlisted personnel which is also a factor in using this technique.

B. Note that our sample size is relatively low. Our sample size is related to the manner in which we collect the data - primarily by team (our personnel) administration. In-house studies have demonstrated that this method yields the highest percent of usable returns, is faster than mail, but is harder on the travel budget.

Thus, we can save money by reducing the sample size. Years ago we thought the generalist concept of a Navy rating would cause us the most sampling problems, along with type and class ship, and paygrade distribution. Experience had proven that the specialization within ratings which are identified by Navy Enlisted Classification codes is the main determinant of sample size. In several instances given a 40% or more usable sample we split the sample taking every other case. The data remained stable to a very low percentage, if NEC's were not considered - even using a structured sample - NEC fallout was unreliable below 20% sampling for the ratings observed.

We have made some significant contributions in NEC validation so it is important that they be adequately covered in our overall sample.

All services except the Army currently use the CODAP programs which were developed by the Air Force some years ago at considerable cost.

We know that the programs of all services generate internal correspondence on the results of their surveys, but the Air Force and the Navy publish comprehensive reports on occupational areas with a distribution list.

There are certain advantages in using a combined question and answer booklet, but on balance, to date we are satisfied with our optical scanned booklets, but this is under constant evaluation.

PART II

SLIDE 2. (See Figure 7.)

All services use task statements, you might say it is a marriage of convenience between task statements and the computer. The arrangement of the tasks may be different - our task statements are randomly inserted in the inventory while the Air Force's are logically structured under certain broad duty categories.

We are going to conduct some studies using the split test and test, re-test approaches in an attempt to ascertain if either arrangement gives us better data - perhaps I should say, different, not better.

Collateral duties are mostly unique to the Navy - these are military duties such as battle stations, ship steersman and various emergency procedures which are not related to specific ratings. We have isolated another group that we designate as "Miscellaneous". It is always "When Occurring" and somewhat rating related - such as "Safety Petty Officer".

All the services collect data on equipment and job satisfaction although there are some small differences in the techniques used, i.e., descriptors and scales. Job characteristics are concerned with physical and mental requirements plus certain broad personality traits.

All services gather data on a relative time spent per task, that is, in relation to all tasks in which the respondee indicates he is involved. Years ago we tried a frequency approach based on absolute time but the results were a disaster.

Note that the Navy is the only service that requires the respondee to indicate the nature of his involvement with the task - formerly we had six items, but now we have dropped the "In-training", and "Instructing" and have four items:

"ASSIST"

"DO" (PERFORM)

"DO AND SUPERVISE"

"ONLY SUPERVISE"

The Navy training community finds that the use of CODAP job descriptions by paygrade and time spent performing coupled with involvement data by paygrades gives an accurate indication of what point in career progression specific training should be given. The involvement data printout is a further refinement also useful in developing the occupational standards mentioned previously and various paygrade correlation studies.

N

O

T

A

P

NAVY OCCUPATIONAL

TASK ANALYSIS PROGRAM

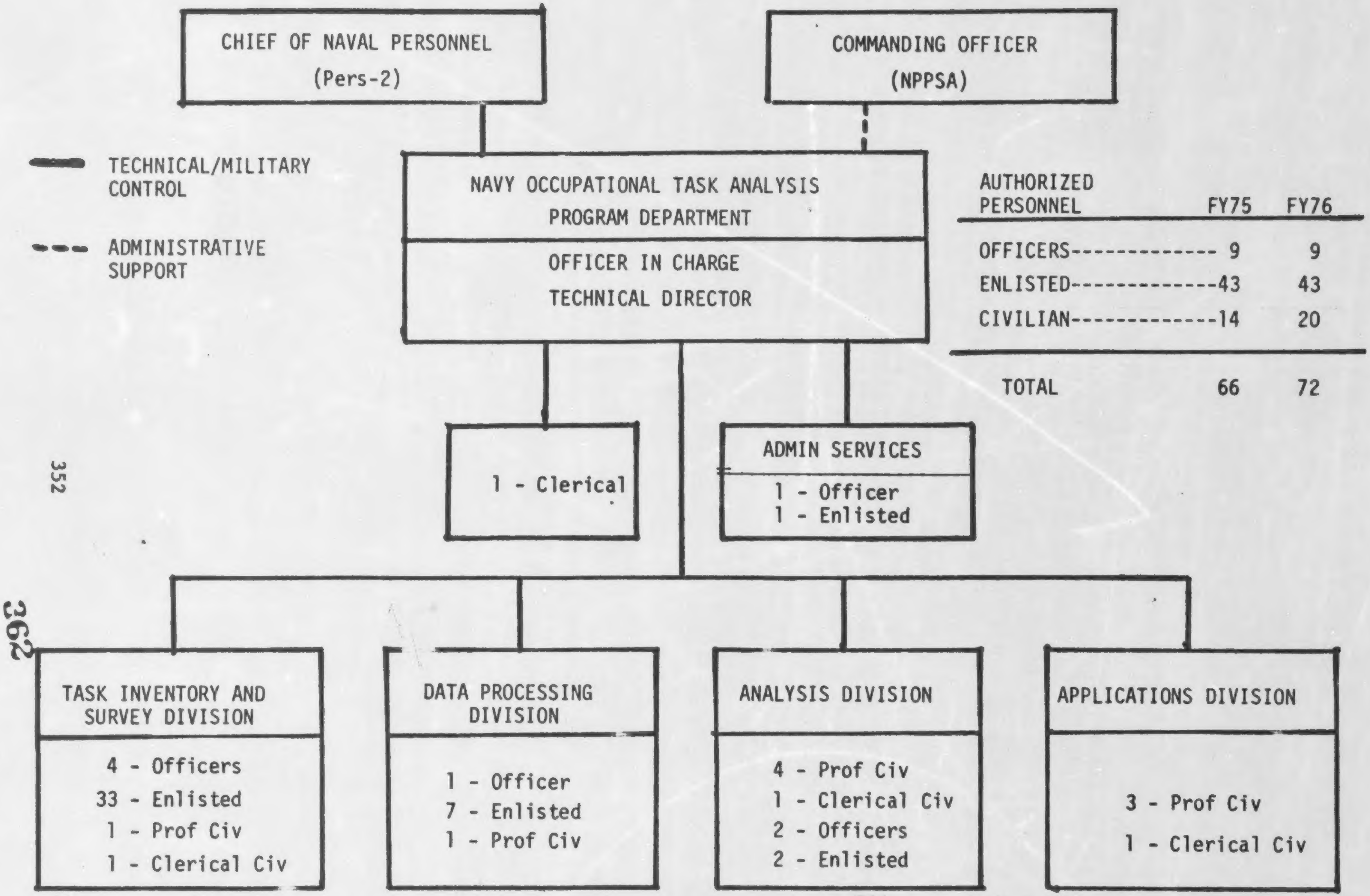
Figure 1

NAVY OCCUPATIONAL TASK ANALYSIS PROGRAM

MISSION

CONDUCT THE NAVY OCCUPATIONAL TASK ANALYSIS PROGRAM FOR THE PURPOSE OF GATHERING OCCUPATIONAL DATA FOR MULTI-PURPOSE USES IN NAVY MANPOWER MANAGEMENT, PERSONNEL ADMINISTRATION, AND TRAINING. CONDUCT COMPREHENSIVE OCCUPATIONAL SURVEYS: MAINTAIN AND UPDATE THE OCCUPATIONAL DATA BANK: ANALYZE OCCUPATIONAL DATA: AND, BASED ON THIS ANALYSIS, SUBMIT RECOMMENDATIONS TO APPROPRIATE AUTHORITIES FOR ACTION. CONDUCT STUDIES TO EVALUATE THE EFFECTIVENESS OF SERVICES PROVIDED TO TASK ANALYSIS CONSUMERS.

Figure 2


CHIEF OF NAVAL PERSONNEL
(Pers-2)
COMMANDING OFFICER
(NPPSA)
TECHNICAL/MILITARY CONTROL
ADMINISTRATIVE SUPPORT
NAVY OCCUPATIONAL TASK ANALYSIS PROGRAM DEPARTMENT
OFFICER IN CHARGE
TECHNICAL DIRECTOR
1 - Clerical
ADMIN SERVICES
1 - Officer
1 - Enlisted
TASK INVENTORY AND SURVEY DIVISION
4 - Officers
33 - Enlisted
1 - Prof Civ
1 - Clerical Civ
DATA PROCESSING DIVISION
1 - Officer
7 - Enlisted
1 - Prof Civ
ANALYSIS DIVISION
4 - Prof Civ
1 - Clerical Civ
2 - Officers
2 - Enlisted
APPLICATIONS DIVISION
3 - Prof Civ
1 - Clerical Civ

AUTHORIZED PERSONNEL	FY75	FY76
OFFICERS	9	9
ENLISTED	43	43
CIVILIAN	14	20
TOTAL	66	72

Figure 3

NAVY OCCUPATIONAL TASK ANALYSIS PROGRAM (NOTAP)

PHASE I PLANNING	PHASE II DATA COLLECTION	PHASE III DATA PROCESSING	PHASE IV DATA ANALYSIS
DETERMINE RATING(S) TO BE ANALYZED	STUDY LITERATURE	SCREEN AND CLEAN RESPONSE PACKETS	CONDUCT IN-DEPT ANALYSIS OF RATINGS
DETERMINE SAMPLE FOR OBSERVATION/INTERVIEW AND SURVEY ADMINISTRATION	OBSERVE AND INTERVIEW JOB INCUMBENTS	OPSCAN RESPONSE PACKETS TO MAGNETIC TAPES	REPORT ON FINDINGS AND RECOMMENDATIONS
INFORM COMMANDS TO BE SAMPLED	COMPILE TASKS IN LOGICAL SEQUENCE	INPUT DATA TO COMPUTER USING CODAP/370 TO PRODUCE STANDARD PRINTOUTS FOR ANALYSIS	CONDUCT SPECIAL STUDIES AS REQUESTED
	CONDUCT TECHNICAL REVIEW OF TASK LISTS	USE CODAP/370 TO PRODUCE SPECIALIZED PRINTOUTS FOR ANALYSIS OR CONSUMER USE	ASSIST CONSUMERS IN DATA INTERPRETATION AS REQUIRED
	PREPARE AND PRINT FINAL JOB TASK INVENTORY (JTI)		
	ADMINISTER JTI'S		
	RETURN OF RESPONSE PACKETS TO NOTAP		

Figure 4

CURRENT IMPACT OF NOTAP ANALYSES
AND DATA

- RATING STRUCTURE
- ENLISTED CLASSIFICATION STRUCTURE (NEC)
- PERSONNEL UTILIZATION
- TRAINING
- QUALIFICATIONS FOR ADVANCEMENT

Figure 5

OCCUPATIONAL ANALYSIS

PROCEDURES AND TECHNIQUES

	AIR FORCE	ARMY	MARINE CORPS	NAVY
OBSERVATION & INTERVIEW	NO	NO	YES	YES
TARGET SAMPLE SIZE	100% for Pop. of 4000 or less	Statistically Valid	50%	30% +
DATA COLLECTION	Mail	Mail	Team Admin.	Team Admin. & Mail
CODAP	YES	PLANNED	YES	YES
PUBLISHED ANALYSIS REPORTS	YES	NO	NO	YES
STANDARDIZED SEPARATE RESPONSE BOOKLETS	NO	NO	YES	YES

Figure 6

(SLIDE 1)

OCCUPATIONAL DATA BANK

	AIR FORCE	ARMY	MARINE CORPS	NAVY
OCCUPATIONAL TASKS	YES	YES	YES	YES
COLLATERAL DUTIES	NO	NO	NO	YES
MISCELLANEOUS DUTIES	NO	YES	NO	YES
EQUIPMENT UTILIZATION	YES	YES	YES	YES
JOB SATISFACTION	YES	YES	YES	YES
JOB CHARACTERISTICS	NO	OCCASIONALLY	YES	YES
RATING SCALES: TIME	9	7	7	5
FREQUENCY	NO	NO	NO	NO
INVOLVEMENT	NO	NO	NO	4

Figure 7

(SLIDE 2)

OCCUPATIONAL ANALYSIS IN THE ROYAL NAVY

By

CDR Colin J. Nicoll

The Royal Naval Manpower Utilization Unit which is staffed by 5 officers and 12 senior ratings is responsible for the conduct of occupational surveys so as to provide analyzed data as a basis for management and training design policy decisions effecting naval manpower including Royal Marines.

The NMUU has been involved in job or task analysis for only some 4 years, but thanks to the wealth of advice and real practical help given by those with a fund of experience in this work on this side of the Atlantic, we are now, I believe, at a similar stage of development of analysis techniques including the use of CODAP. Even so, it is flattering to be asked to contribute at so august a gathering when I am such a new member of the club.

I have been asked to give a run down on our programme which I will do by first outlining some highlights of results from our recent surveys including our first CODAP-based study. I will go on to describe the surveys of the Royal Naval Engineering branches at present being administered and also a recent officer survey in which we introduced some new ideas. Finally, I will look ahead into the future of job analysis not only in the Royal Navy but also in other British services.

The main purpose of our studies is to provide job descriptions for use as a basis for training design but we have found increasingly that we can provide valuable data for other manpower policy decisions. This is a list of our 1974 studies and I will mention some of the results as examples of how we contribute to training design and sometimes go further.

Writer ratings are the clerical branch responsible also for pay and cash. There are both male and WRNS ratings. Amongst other things, the survey showed that there is a lot of time spent on using and maintaining office machinery for which there was no training, that the tasks of certain senior ratings differed considerably whether they were serving ashore or at sea underlining the need for a refresher course and that a proportion of the WRNS, although half their training syllabus is dedicated to work with service documents, spend 80% of their time typing and the other 20% making tea.

Still talking about WRNS, the Cinema Operators Branch, originally formed before the general introduction of simple 16mm film projectors, were becoming sad and unwanted. Our study directive was to seek out suitable alternative employments with some promise of job enrichment. It has now been agreed that the girls be responsible for all types of training aids from Vu-graphs to closed circuit television and be renamed education assistants.

We made a small study into the work of radio electrical mechanics in submarines. This study showed that 50% of the senior and 70% of the junior ratings' time is spent in their submarines on work quite unconnected with their technical training. This does not indicate misemployment but showed up a weakness in the PR of their early training which previously did not tell them what to expect: first a submariner and only then a radio mechanic.

We made a very detailed task analysis into the professional duties of radarmen working in the operations room purely for training design.

The Air Engineer Officer study was conducted by an outside authority with NMUU advice and gave valuable information at a time when the Royal Navy is phasing out fixed-wing carriers and going all rotary wing.

In the Royal Navy some of our frigates have a small detachment of Royal Marines--about 24 in all. These replace seamen and are required to carry out seamanlike tasks for which they are largely untrained. I am glad to say that as a result of our study, this situation is being changed.

Last year, we conducted a trial analysis of the RN communications branches using CODAP. The object of the trial was to confirm the suitability of the program for RN analyses and I am glad to say that this was amply proved. The communications branch was chosen because it has three clearly structured sub-branches and is of manageable size for a trial analysis--about 3,000 men.

We broke down the inventory into the usual sections: background identifiers both of service and private background, including marital status, whether they owned a home and so on; job content which was the main breakdown of tasks to be measured by time involvement and level of involvement; two sections on the use of equipment and publications; a section asking for worker characteristics required, such as skill of hand, strength or arm or mental stamina; some questions on the ratings' opinions about specific matters concerned with conditions of service; and finally, general comments. I should say that we will not use either worker characteristics or general comments again; the first because we were given blinding glimpses of the obvious, the second because I believe it is not my business to be an avenue of protest, but to analyze a valid sample of fact or opinion even though we had some intriguing remarks *such as "Couldn't this be filled in by my wife?"*

Some results of the survey of interest were that we showed that the use of hand key Morse at sea although much reduced by the installation of teleprinters is more than was subjectively believed. The Signal School has had to introduce a "save Morse" campaign to ensure there are enough competent operators at sea. We were able to show that the distribution of technical books can bear considerable revision. Some are seldom consulted even though they provide guidance on day-to-day affairs. They perhaps need rewriting--or perhaps they can be done away with altogether.

I would like to turn now to the biggest study which we have attempted. This is a CODAP survey of the entire Engineering Branches of the Royal Navy. The Ministry of Defence in London is currently conducting a fundamental appraisal of the structure of all Naval engineers--air, surface, submarine, marine electrical, weapons, radio: the lot. The study team will report to the Admiralty Board in the new year their broad recommendations--to take two examples, should the responsibility for the generation and distribution of high power electricity remain with the weapons electrical branch or be transferred to the Marine Engineers--and do we need as many highly trained, skilled craftsmen at sea when the normal method of repair is by replacement.

The NMUU has been directed to conduct a survey in support of this study with the aims of quantifying work content, determining craft skills used at sea, assessing ratings' opinions on service life and identifying factors which may influence job satisfaction.

To do this, we will administer 6 differing questionnaires to the fleet and shore establishments starting this week. This shows the magnitude of our endeavor. The tasks are those performed by a third of the Navy.

The contents of each inventory are similar to those for the communications trial except that we have left out worker characteristics and comments and included some questions on job satisfaction.

We realize that the information to be gleaned from the ratings who complete the questionnaires is correctly limited to actual involvement and that there are other things of equal or greater weight in deciding branch structures and training design. We are, therefore, arranging to collect what I describe as managerial inputs. From ships officers we are asking for the assessment of the minimum grade of ratings who should do each task. This may well differ from the man who perforce does it. From our sea training command we are going to be told those tasks which are habitually badly performed during the work up period and from experienced trainers we are going to try to assess the difficulty of each task by asking their assessment measured as relative time to learn.

We hope to have our data bank available for access by the middle of January and then be ready to provide our customers with the hard statistics they want. There is however the lack of a customers guide to CODAP, as opposed to a users guide, which would save me a lot of the sales talk and explanation which is at present necessary.

The last study I would like to mention is that of the junior supply officer and WRNS secretarial officers which is about finished. We asked these officers more than just their time involvement in tasks. At a more detailed level than that which we considered necessary for time, we asked for assessments of difficulty, whether the officer was monitored in or delegated the task, what type of training he or she considered to be necessary, and whether he would have liked more training.

At the same time we have sought a managerial input from senior supply officers, in this case giving them bench marks of importance. Although the purist might prefer a sliding scale, we considered for this initial attempt at correlating both the senior and junior officers' opinions, we should give as much guidance as possible.

We have not fully analyzed the importance input, but it seems that there will be some interesting results--the junior officers would like to delegate and consider they need little training in some tasks which their superiors assess to be of prime importance.

An interesting finding of the survey has been that we have shown that from the pure work content standpoint, the training and shore appointments of the male and female officers of the specializations could be common. There may well be many overriding considerations but our report has ruffled the surface of traditional male prejudice.

And so to the future. Apart from the major work of completing and analyzing the engineers' study, we will be conducting more CODAP surveys though probably of less ambitious scope. I am not yet certain how far up the ladder of seniority of officers our work is appropriate. I feel that once the job becomes supervisory or managerial rather than actually doing, there are better ways of assessing the requirement for the officer. Similarly, I am in doubt over analyzing the combat arms. If you analyze a man in a training situation, all you will do is compound the shortcomings of your current training.

The other services have shown an interest in our work, particularly the Army and I now expect to see them using CODAP to survey either the Royal Electrical and Mechanical Engineers or the Royal Signals in the first place.

I hope I have put across that occupational analysis is now accepted in the Royal Navy as a valuable means to improving our use of men in times when they are hard to recruit, hard to retain, and yet required to operate and maintain highly complex equipment. We must have sound management and clear objective training if we are to survive as a competent fighting service.

THE ROYAL AUSTRALIAN AIR FORCE OCCUPATIONAL MEASUREMENT PROGRAM

PAST, PRESENT, AND FUTURE

by

Major Wayne S. Sellman
Directorate of Personnel Plans
HQ United States Air Force
Washington, D.C.

Squadron Leader John W. K. Fugill and Captain Gary R. Lawrence
Headquarters Support Command
Royal Australian Air Force
Melbourne, Victoria

INTRODUCTION

The Royal Australian Air Force (RAAF) occupational measurement program was established in 1967 following an overseas review of techniques used by the Royal Air Force, Canadian Defense Forces, and United States Air Force (USAF). After several modest efforts using a combination of methodologies, the RAAF sought closer co-operation with the USAF through the formulation of an exchange program in the occupational analysis area. This exchange became a reality in June 1970 with the assignment of a RAAF squadron leader (major) education officer to the Air Force Human Resources Laboratory in San Antonio to study occupational research techniques and operational job analysis procedures. At the same time, the USAF exchange officer, a captain psychologist, was posted to Headquarters Support Command in Melbourne to develop and supervise the implementation of appropriate job analysis and training evaluation procedures within the RAAF. Thus, the RAAF system of occupational analysis which has evolved over the past five years is a modified version of the USAF job inventory approach pioneered by the Occupational Research Division of the Air Force Human Resources Laboratory.

Obviously, there is no way we can discuss all aspects of the RAAF occupational measurement program in the limited time we have today. What we would like to do is make some observations about the techniques the RAAF uses to collect, analyze, and report job data, some of its practical applications, and where we hope to go from here.

RAAF OCCUPATIONAL SURVEY PROCEDURES

Although the job inventory approach to occupational analysis (Morsh, 1969; Morsh, 1964; Morsh, Madden, and Christal, 1961) is well known to most of you, we would like to briefly summarize the RAAF version. First there is the construction of the job inventory. As you know, the job inventory is composed of two basic sections: background information questions, and a list of the significant tasks performed by airmen at all levels of rank and job proficiency within a mustering (USAF specialty). An initial listing of tasks is constructed from available published sources such as manuals, regulations, training documents, etc. This list is then subjected to an interview review by senior NCO subject matter specialists, and to a mail review by additional senior NCO specialists working in the field. The preliminary listing of tasks is then revised in light of their comments and recommendations for changes, additions, or deletions, and the final version of the job inventory is prepared.

The inventories are then mailed to unit trade test officers (analogous to USAF test control officers) who administer them under standardized conditions to selected groups of job incumbents in base testing rooms. We should note here that great care is taken to select representative samples. In the larger musterings ($N>800$) samples between 50 and 75 percent are used while 100 percent sampling is the goal for the smaller musterings.

In completing the inventories, job incumbents provide identification, background, and other job related information. Then using specially designed machine-readable answer sheets, they respond to the task statements by indicating each task performed and the relative amount of time spent on it. Finally, incumbents add tasks not listed in the inventory. Because the task response answer sheets used are to be optically scanned, incumbents write-in their additional tasks in the background information section, not on the task answer sheets themselves.

Completed inventories are returned to us at Headquarters Support Command where they are checked for errors. Background information is forwarded to the RAAF EDP Centre for keypunching and transfer to magnetic tape. The task responses go to a local university where they are converted to tape by an optical scanner. (By the way, the optical scan error rate for our last 25 surveys was less than $\frac{1}{2}$ percent.) The task response tape is then sent to the RAAF EDP Centre where, by means of subject respondent codes, the two tapes are merged into one which contains the complete file for each airman in sequence. The data are then re-edited and prepared for computer analysis.

One of the major advantages of the job inventory approach to occupational analysis is that the results are comprehensive and easily quantified (Christal, 1971). To fully exploit this advantage, we are working with the RAAF EDP Centre in the development of a catalogue of standard computerized occupational analysis programs. To date, we have available two programs for application to our job surveys - group job descriptions, and group difference descriptions. The group job description program enables the computer to generate job descriptions, made up of tasks performed and the average percentage of time spent on each of them for any group of individuals where the groups can be defined in terms of background variables. For example, a job description can be computed for all airmen of a given rank, or for those assigned to a specific work area, or for those who have been on the job for less than six months. Indeed, a group job description can be generated in terms of any single or combination of background variables. Thus, a group job description could be provided of all corporal, electrical fitters, between the ages of 20-25, with over three years experience in the mustering, and over one year in their present position. The background variables and possible combinations on which groups may be categorized are limited only by the job analyst's imagination.

Our second program provides comparisons between groups. These group difference descriptions effectively point out the difference between two groups in terms of the emphasis they place upon various tasks. With such descriptions, it is possible to determine how the time spent on tasks varies with rank, time in the job, time in service, by unit, by area of assignment, or any other background variable completed by the incumbents. For example, a comparison by rank would depict how the job changes as one progresses in rank; or which tasks are performed by junior members and which by more senior ones. A comparison of major work locations would indicate which tasks are performed by airmen assigned to flying squadrons as opposed to those which are performed by aircraft depot members.

Those of you familiar with the job inventory approach will immediately recognize that the above described programs are versions of two CODAP (Comprehensive Occupational Data Analysis Programs) programs designed by Dr Raymond E. Christal of the Air Force Human Resources Laboratory.

The RAAF has recently acquired the complete CODAP system from the USAF and will soon have it in operation. In this regard, we will join the U.S. Marine Corps, U.S. Navy, U.S. Army, U.S. Coast Guard, and Canadian Defense Forces in taking advantage of this, the latest and most powerful in computerized occupational analysis methodology. Greater detail concerning the specifics of CODAP can be found in the writings of Archer (1966); Morsh and Christal (1966); Morsh (1969) and Christal (1971, 1973).

SOME APPLICATIONS OF OCCUPATIONAL INFORMATION

So far we have described how occupational information is collected and analyzed. What you may want to hear now are some experiences in using the data. Actually, occupational data can and should impact on almost every part of the personnel system. We will, however, only have time to mention a few of the applications being implemented by the RAAF.

Design of Training Systems

One valuable service to be rendered to the training system is to ensure that course content is relevant to the work that airmen will have to perform upon graduation. On the question of task selection criteria, there has always been uncertainty as to the factors to be used and the weights to be applied to those factors. A review of current practices in Australia and the United States revealed that many so called task selection factors (i.e., importance, criticality, complexity, etc.,) were clearly mult-dimensional and so subjectively interpretable that work supervisors were unable to give reliable information about them.

Using insights provided by Christal (1970), we have defined a set of work - oriented factors on which RAAF supervisors are able to give fairly reliable information (Fugill and Sellman, 1974). The factors are: Task difficulty, frequency of inadequate performance, consequence of inadequate performance, probability of emergency performance, and skill perishability. A simple unit weight algorithm was then developed using factor data supplied by a group of senior work supervisors (Fugill and Sellman, 1974; Sellman and Fugill, 1974). The algorithm for translating job information into training requirements thus took the following form.

Tasks Performed by 10% or More Workers	Minus	Non-Trainable Tasks	Equals	Relevant Trainable Tasks
Relevant Trainable Tasks	Minus	Low Task Difficulty or Low Consequences of Inadequate Performance	Equals	Work - Oriented Graduation Requirement

To the work -oriented graduation requirement, training specialists add those knowledges and skills which are required to maintain newly acquired equipment.

Job Satisfaction

As more and more emphasis is placed on improved personnel utilization within an environment of defense strength and budget reductions, job satisfaction has begun to take on an increasing importance within the military services. Following a methodology devised by the USAF, in 1973 the RAAF began to capture job satisfaction data as part of the regular occupational survey procedures (Sellman and Fugill, 1973). Respondents to job inventories use a 7-point rating scale to indicate their job interest ranging from "extremely dull" to "extremely interesting" and the extent to which their jobs use their talents and training with responses from "not at all" to "perfectly."

To date, data have been collected for over sixty percent of the RAAF's airmen and airwomen in approximately 35 career areas (musterings). Detailed analyses of data are currently underway, but a few observations are now possible. Differences in expressed job satisfaction have been found to exist between musterings and among individuals within musterings. Table 1 shows, for example, the rank - ordered positions of 13 musterings on the three satisfaction factors (scores standardized).

TABLE 1

INDICES OF JOB SATISFACTION
MUSTERINGS RANK - ORDERED FROM HIGH TO LOW

Job Interest	Utilization of Talents	Utilization of Training
Radio Technician (Ground)	Clerk	Telecommunications Technician
Engine Fitter	Clerk (General)	Clerk Supply
Supplier	Engine Fitter	Service Police
Senior Clerk Supply	Radio Technician (Ground)	Engine Fitter
Telecommunications Technician	Telecommunications Technician	Clerk (General)
Clerk (General)	Supplier	Radio Technician (Ground)
Aircraft Metal Worker	Service Police	Clerk
Steward	Clerk Supply	Supplier
Service Police	Security Guard	Security Guard
Clerk (Admin)	Steward	Steward
Clerk Supply	Aircraft Metal Worker	Clerk (Admin)
Security Guard	Clerk (Admin)	Aircraft Metal Worker
Clerk	Senior Clerk Supply	Senior Clerk Supply

This information allows management to identify the musterings whose members report relatively high and relatively low "satisfaction." The positioning of musterings at the bottom of the lists suggest possible problems in the way in which their work is structured. Management review is indicated if problems inherent in the work are to be identified and solutions achieved.

The long-term job satisfaction research program of the RAAF is directed so personnel managers will be able to examine systematically the following types of problems.

a. What is the expressed job satisfaction of airmen who decide not to reenlist?

b. How can job satisfaction indices be used to predict reenlistment rates?

c. What are the characteristics of individuals within defined job types who express high - low job satisfaction?

d. What is the relationship between rated job effectiveness (airman performance reports) and expressed job satisfaction?

e. What is the trade-off between job satisfaction, aptitude, education, training, and work experience?

f. How can jobs be modified so that workers will look favorably on the RAAF as a career choice?

Career Planning

With this application, the aim is to guide airmen through a planned series of job to make best use of their talents and experience, and to enhance their job satisfaction. Since a job inventory contains all of the significant tasks being performed by airmen in a mustering, it is an ideal framework for establishing and maintaining individual experience records. CODAP programs can prepare the records in a format similar to the one shown in Table 2.

TABLE 2

INDIVIDUAL EXPERIENCE RECORD
TASK LEVEL

Control Number: 046
Rank: LAC
Name: Burkett, J. R.
Total PAF Service: 36
Time in Present Rank: 22
Time in Present Job: 10
Number of Subordinates: 0

Job Title: Voucher Raiser
Job Interest: 3
Util of Talents: 4
Util of Training: 3
Service Number: A53421
Mustering: CLKSPLY
Unit: 100 Sqn

Duty/Task	Task	Percent Time Spent	Cumulative Percent
R 427	Raise countership vouchers	27	27
R 464	Raise transfer vouchers	25	52
R 435	Raise exchange vouchers	17	69
G 156	Action AOG, EOS, URR and VOR demands	16	85
G 162	Hasten overdue demands	5	90
G 162	Hasten overdue demands	5	95
G 162	Hasten overdue demands	5	100

Over a 20-year period, an airman's personnel folder might accumulate six to ten experience records. With this detailed information, a set of posting options could be exercised to create "desirable" career progression.

When an airman arrives on posting, some commanders tend to optimize his past experience by placing him in a work position similar to the one he has just vacated. Although this makes the airman immediately useful (and minimizes the need for on-the-job-training), it is often detrimental to the airman's professional development. One way to reduce this may be to implement, in peace-time, a deliberate posting policy of mis-matching.

SPECULATIONS ON THE FUTURE

In December 1973, the RAAF Department of Air was abolished as an operating entity and was integrated along with the Department of Army and Navy into the Department of Defense. Under the new reorganization the RAAF retained its separate identity as a service, but its planning and administrative functions were centralized within the expanded Department of Defense.

There is presently some discussion of combining trade groups and training courses across the three services and of developing a common defense monitored tri-service occupational analysis system. Should such amalgamation come to pass, well researched methodologies would be essential. Since CODAP is the only comprehensively researched system in existence, the case for the adoption of the RAAF job survey system is a strong one.

From a Department of Defense viewpoint, the advantages of using CODAP are obvious. First, some cost-savings would be achieved by maintaining a single set of multi-purpose computer programs to support the three military services' occupational survey programs. Secondly, the outputs would have common formats which would make each service's data more easily interpretable by the other two. Finally, after determining the similarities (or differences) of the various trades across the services, substantial cost-savings could be effected by combining training courses and sharing work facilities.

SUMMARY

Occupational analysis in the Royal Australian Air Force has developed over the past seven years into a highly sophisticated methodology designed to meet the needs of a modern Air Force. The feasibility of the job inventory approach has been demonstrated, and we are constantly striving to develop innovations and advances in our inventory construction and survey administration techniques. Clearly, however, the progress that has been made in these directions does not lead to the end of the road. It is very encouraging to note that within a span of only a few years, we have been able to talk about not only what will take place in the future but also about past accomplishments and what is occurring now. We believe that the time is right to show that the RAAF's occupational measurement program is the best in Australia. With the acquisition of CODAP, RAAF leadership is a fait accompli.

REFERENCES

Archer, W. B. Computation of group job descriptions from occupational survey data. PRL-TR-66-12. Lackland AFB, TX: Personnel Research Laboratory, Aerospace Medical Division, December 1966.

Christal, R. E. Implications of Air Force occupational research for curriculum design. Paper presented to the Minnesota Research Co-ordinating Unit for Vocational Education Seminar: Process and Techniques of Vocational Curriculum Development. University of Minnesota, April 1970.

Christal, R. E. Combining occupational information with personnel data to resolve manpower management problems. Paper presented to the NATO Conference on Manpower Planning Models, Cambridge, England, September 1971.

Christal, R. E. The United States Air Force Occupational Research Project. Paper presented in Symposium, the State-of-the-Art in Occupational Research and Development, Navy Personnel Research and Development Center, San Diego, California, July 1973.

Fugill, J. W. K. and Sellman, W. S. Design of work - oriented task selection algorithms for the development of vocational training courses. Paper presented to the 9th Annual Conference of the Australian Psychological Society, Perth, Western Australia, Australia, August 1974.

Morsh, J. E. Job analysis in the United States Air Force. Personnel Psychology, 1964, 37, 7-17.

Morsh, J. E. Collecting, analysing, and reporting information describing jobs in the United States Air Force. Paper presented to the 77th Annual Convention of the American Psychological Association, Washington, DC, September 1969.

Morsh, J. E. and Christal, R. E. Impact of the computer on job analysis in the United States Air Force. PRL-TR-66-19. Lackland AFB, TX: Personnel Research Laboratory, Aerospace Medical Division, October 1966.

Morsh, J. E., Madden, J. M., and Christal, R. E. Job analysis in the United States Air Force. WADD-TR-61-113. Lackland AFB, TX: Personnel Laboratory, Wright Air Development Division, February 1961.

Sellman, W. S. and Fugill, J. W. K. Occupational analysis in the Royal Australian Air Force. Paper presented to the 15th Annual Conference of the Military Testing Association, San Antonio, Texas, October 1973.

Sellman, W. S. and Fugill, J. W. K. Development and application of an algorithm for translating job information into technical training requirements. Paper presented to the 16th Annual Conference of the Military Testing Association, Oklahoma City, Oklahoma, October 1974.

SECTION X

USING TEST RESULTS IN THE TRAINING PROCESS

TEST RESULTS

THE KEY TO THE MANAGEMENT OF TRAINING

By

Robert N. Johnson and MAJ Richard F. Malone
U. S. Army Institute of Administration

With the ever increasing reductions being made in the resources available for training and the never decreasing demands for quality output from the training system, training managers are faced with perhaps their greatest challenge: How can we get the best possible product for the least cost? What indicators are available to tell us where we should concentrate our resources? In the U. S. Army Institute of Administration, we believe that we have developed a system that provides these indicators and assists the manager in allocating his resources. These indicators are the outputs of our testing system - test results - and are readily available to every instructor and instructional manager in the Institute. The purpose of this presentation is to describe how we design performance tests to provide the necessary output and how we use these outputs to conserve resources and promote cost effective quality training.

As a portion of our course design process, we perform a detailed analysis of each task to be trained, identifying each behavior required to accomplish the task including all variations of the task which might actually occur on the job in the field. After the analysis, we design the performance test, prior to and independent of instructional design. The test always comes first. Our performance tests attempt to realistically approximate the real world to include identical cues and conditions and provide the student with the same options which he would find on the job. The scoring of the test must be objective - each element of the performance is either done or not done, right or wrong. To aid in achieving objectivity, we design the score sheet first. Then the materials which will be required to perform the task are developed. The score sheet usually takes on one of three forms:

a. The score sheet may be a checklist for the instructor to use in evaluating a product which the student has produced, e.g., a typed letter, a processed mail parcel, a completed form, etc..

b. The score sheet may also be a checklist which an instructor uses as he observes each student perform the required steps in a process which does not result in an observable and measurable product, - e.g., correcting a machine stoppage on a 188 collator.

c. An answer sheet on which the student records his decisions with respect to the key mental decisions involved in a primarily cognitive task - e.g., review of outgoing correspondence prior to signature.

Regardless of the form the score sheet takes, each item represents an element of performance of a real task in terms of the characteristics of an acceptable product, steps in an acceptable process or a specified number of variations of a task or its conditions. Please bear in mind that the use of the term "item" in this paper refers to an element of performance within a single task - rather than a representative sampling of independent multiple choice items selected from a grouping of tasks.

For all types of performance tests, the results of each administration are summarized by our computer in terms of a frequency distribution of unweighted scores with a single line summary for each class as shown in Figure 1.

```
                         DISTRIBUTION OF SCORES

RAW       T       NUMBER OF STUDENTS

56        U       14 XXXXXXXXXXXXXX
55        U        9 XXXXXXXX
54        U        6 XXXXXX
53        M        5 XXXXX
52        M        5 XXXXX
51        L        1 X
50        L        1 X
49        F
48        F        1 X
47        F
46        F

                           COMPARATIVE HISTORY
CLASS     STU                CLASS SCORES                PERCENT OF STUDENTS
                  MAX           PASS     AVG      FAIL   L-3    M-3     U-3
75-03     42      56             50      55.1      2%     6%    24%     68%
```

Figure 1

As you can see the frequency distribution displays overall class performance in descending order of achievement. On this specific test 56 elements of performance of a single product or process were evaluated. Of the 42 students in the class, 14 achieved a perfect score, 9 achieved 55, 6 achieved 54, etc.. The X's provide easy visualization of the spread in scores. Since the criterion for this test is 50 out of 56, the computer identifies any lower score with an F for failure. The passing scores 50 thru 56 are then divided into lower (L), middle (M) and upper (U) thirds of the passing range. It is this division of class scores into failures and thirds of the passing range that permits us to summarize current and past overall class performances for management purposes.

Note the single line of data under the heading Comparative History pertaining to this class #75-03. This data is nothing more than a summarization of the same data displayed in the frequency distribution above. Forty two students took the test; the maximum attainable score was 56; the criterion was 50 and the class average was 55.1. The frequency of scores has been compacted into the percentage of students who failed (fail), and those who attained each of the thirds of the passing range (L-3, M-3, U-3).

In addition our computer program displays under the Comparative History section the performance data for the last 9 administrations of the test. For illustrative purposes however, Figure 2 shows the last 14 administration of the same instrument.

COMPARATIVE HISTORY

CLASS	STU	CLASS SCORES			PERCENT OF STUDENTS			
		MAX	PASS	AVG	FAIL	L-3	M-3	U-3
75-02	40	56	50	55.0		3%	30%	67%
75-01	39	56	48	54.4		7%	33%	60%
74-11	40	56	48	53.2		10%	37%	53%
74-10	42	56	48	54.3		4%	33%	63%
74-09	44	56	48	53.7	3%	18%	24%	55%
74-08	37	56	45	52.8		19%	34%	47%
74-07	43	56	45	52.9		12%	39%	49%
74-06	39	56	45	51.4	5%	12%	63%	20%
74-05	36	56	45	52.0	2%	10%	57%	31%
74-04	37	56	42	49.0	8%	8%	68%	16%
74-03	35	56	42	47.1	15%	45%	30%	10%
74-02	38	56	42	43.7	18%	51%	29%	2%
74-01	34	56	42	41.9	24%	48%	28%	

Figure 2

It is this display of the history of class performance on a single test or task that allows us to manage our resources. Note that while the last few classes have experienced little difficulty, earlier classes had extreme difficulty with the exact same instrument.

For management purposes we classify the quality of training into three categories: undertraining, overtraining and adequate training. Let us examine the management aspects of each category beginning with undertraining. We define undertraining as a condition which exists when over a period of time (several classes), test results indicate that over 10% of the students fall into the failing (fail) category. Figure 3 displays the undertraining category.

COMPARATIVE HISTORY

CLASS	STU	CLASS SCORES			PERCENT OF STUDENTS			
		MAX	PASS	AVG	FAIL	L-3	M-3	U-3
74-03	35	56	42	47.1	15%	45%	30%	10%
74-02	38	56	42	43.7	18%	51%	29%	2%
74-01	34	56	42	41.9	24%	48%	28%	

Figure 3

We consider the instructional system to be ineffective when 10% or more of the students fail the test on a first attempt. Note that this condition exists in the three classes shown. Assuming a constant level of quality of input we take the position that some aspect of the instructional system must be modified and that it is our job to assist instructors and instructional supervisors in identifying problem areas. Accordingly our computer furnishes for each test a summary of item information in the form of Miss Rates as shown in Figure 4.

ITEM	MISS RATE
1	21%***
2	7%
3	9%
4	1%
5	9%
6	
7	7%
8	48%***
9	15%
10	46%***
11	42%***
12	7%
13	34%***
14	
15	55%***
16	46%***

Figure 4

Since each item represents an element of performance of a real task in terms of characteristics of an acceptable product, steps in an acceptable process or a specified number of variations of a task or condition, the higher Miss Rates pin-point areas of instruction which require additional emphasis or modification. Miss Rates in excess of 20% are flagged with three asterisks to alert supervisors of major problem areas.

As these problem areas are identified and eliminated, Miss Rates in future classes fall, resulting in higher individual and class average scores and the class distribution will no longer indicate undertraining. In most cases, the mere identification of the problem area is sufficient for instructors or instructional package designers to eliminate the problem. In other cases, however, the committment of additional instructional time or resources is indicated.

As you all know instructional time is at a premium because this time means money. Where can we get more time without expending additional resources? To answer this question, let us look at the category of training we classify as overtraining. We define overtraining as a condition which exists when over a period of time (several classes) test results indicate that over 90% of the students fall into the middle and upper thirds (M-3 + U-3) of the passing range. Figure 5 portrays an overtraining condition.

COMPARATIVE HISTORY

CLASS	STU	CLASS SCORES			PERCENT OF STUDENTS			
		MAX	PASS	AVG	FAIL	L-3	M-3	U-3
75-03	42	56	50	55.1	2%	6%	24%	68%
75-02	40	56	50	55.0		3%	30%	67%
75-01	39	56	48	54.4		7%	33%	60%
74-11	40	56	48	53.2		10%	37%	53%
74-10	42	56	48	54.3		4%	33%	63%

Figure 5

Note that over a period of time, few, if any, students ever fail this test and the majority (at least 90%) of those who pass exceed the criteria by at least one third of the passing range. When this overtraining condition exists we follow the specific course of action shown at Figure 6.

REQUIRED ACTIONS - OVERTRAINING

1. Review validity of test instrument.

2. Review adequacy of test criterion.

3. Arbitrarily reduce instructional resources (instructional time, number of instructors, number of student group, etc.)

Figure 6

---

a. Review of the face validity of a test is a continuing process. If the test does not adequately represent the task and task conditions, it must be revised.

b. In many cases we find that our initial estimates of the criterion were too low. Even today many instructors and managers continue to insist on a 70 or 75% standard without consideration of the actual higher standard required on the job. By raising the standard for future classes we tend to eliminate the test from the overtraining category. If an overtraining condition persists we continuously review the criterion and keep raising it until we are confident that it is a valid representation of the real world requirement.

c. If the test is valid and if the criterion is considered adequate then we arbitrarily reduce our committment of resources to that specific block of instruction. Since our major resource is instructional time, this is our emphasis. If 14 hours of instructional time are currently conducted, we reduce it to 13 hours forcing our instructors to combine, condense, or eliminate in order to meet time constraints. If future test results do not drop we force further reductions until results no longer indicate an overtraining condition.

As you can see, although we start from the assumption that our task, conditions and standards are valid, our courses of action however, include the possibility that these assumptions may have been erroneous. We reduce resources committments only after reconsideration of the validity of the test and the criterion. What do we do with the instructional time or resources generated? Well, we have several options:

a. Reallocate them to other blocks of instruction which are currently portraying an undertraining condition as shown earlier. In effect we steal from Peter to pay Paul - an age-old approach - but with a new twist - we have evidence that Peter is in fact rich, and Paul is in fact poor in terms of the current quality of student performance. Instructional resources are therefore reallocated based on quality as measured by test results.

b. We can also dedicate the additional time or resources to other tasks which should have been included in the training program but were excluded due to resource constraints.

c. And lastly we can reduce the length of the course - a course of action that is sure to please TRADOC, our senior training Headquarters.

There is, of course, one remaining category of test results, those that indicate neither undertraining nor overtraining. Figure 7 is an example of test results that illustrate this condition.

COMPARATIVE HISTORY

CLASS	STU	CLASS SCORES			PERCENT OF STUDENTS			
		MAX	PASS	AVG	FAIL	L-3	M-3	U-3
75-02	41	27	21	25.7	2%	10%	29%	59%
75-503	31	27	21	25.4	3%	16%	29%	52%
74-16	51	27	21	25.2	2%	16%	35%	47%
74-13	39	27	21	24.8	3%	28%	31%	38%

Figure 7

Note that failure rates average below 10% and that the frequency distribution by thirds of the passing range indicates a reasonable spread with the desired negative skew. It is this category of test that we label as ready for packaging. Our test results show that the current instructional system is achieving the instructional objectives in a reliable manner. The name of the game, now, is to stabilize the instructional system before someone tampers with a good thing. At this point our instructional design personnel work with the instructors to identify and implement methods and media which will produce that same quality at a lower overall resource cost. Programed texts, television, CAI or any other stabilizing media are considered. Self paced exportable packages are, of course, our primary goal. It should be noted that although the delivery system is altered, the proven performance test remains the measure of instructional effectiveness.

To this point I have illustrated our management approach on a test by test or task by task basis. We also are beginning to periodically provide our instructional managers with summarized information which can be compared on a course by course basis as shown in Figure 8.

TEST RESULTS
ENLISTED SKILL PROGRESSION COURSES

COURSE	# TESTS	UNDER-TRAINING	OVER-TRAINING	% ADEQUATE
71C (Steno)	20	0	6	70%
71D (Legal)	24	10	3	46%
71F (Postal)	17	1	2	82%
73C (Finance)	11	4	2	45%
73D (Acctg)	7	2	2	43%
74C (Pers Data Anal)	12	3	1	67%
74D (Mach Op.)	11	0	1	91%
74F (Comp Sys Op)	7	0	4	43%
74F (Comp Prog)	9	2	1	67%
74G (Sys Anal)	6	0	3	50%
75B (Unit Clk)	14	4	4	43%
75C (Pers Mgmt)	31	10	5	52%
75Z (Pers Sr Sgt)	32	11	8	41%
00E (Recruiting)	8	0	5	38%
Institute	209	47	47	55%

Figure 8

By examining this data, managers can compare progress in each course, against each other and against the Institute average. In addition we provide summary cost data in terms of instructor platform hour requirements, density of academic instruction, facility requirement summarizations, overall cost per student and similar information which provide the basis for the necessary cost/gain tradeoffs.

Our approach obviously requires extensive computer support and, for those of you in the Army school system, the Institute of Administration has recommended a major redesign of the test scoring and grading portion of the TRADOC Education Data System (TREDS) to provide the data necessary to produce this type of management information.

Summarized results of valid performance tests are the key to managing the effectiveness and efficiency of our courses at the Institute of Administration. They tell us when there is not enough training, too much training, or just enough. They are the answer to the challenge - get the best possible product for the lowest possible cost!

TASK ANALYSIS BASED TESTING AS A MEASURE OF TRAINING EFFECTIVENESS

LCDR Clarence L. Walker, USN
Officer in Charge
Central Test Site for Personnel and
Training Evaluation Program (PTEP)

The Fleet Ballistic Missile Weapons System Training Program is a training management system whose elements are structured on equipment based task analysis. This analysis uses, as a first block, the skills and knowledges needed to support a system, a subsystem or piece of equipment. These skills and knowledges are then related to an individual through leveling statements which define an individuals responsibility in a particular area. The nomenclature and implementing matrices associated with the program are not pertinent to this paper. The relationship between theoretical knowledge and skills however is.

Figure 1 shows the way levels of theory are used to support skill achievement. T, represents theory; O, operations; P, preventive maintenance and C, represents corrective maintenance. Increasing subscripts show increasing difficulty. In each case theory training exists only to support specific skills. The line below O_1, P_1, C_1 indicates the extent of training provided at replacement training; other training is done either at advanced training sites or onboard ship.

The Central Test Site for the Personnel and Training Evaluation Program, (PTEP) is tasked with measuring the effectiveness of the FBM weapons system training program in meeting the needs of the fleet. In doing this it measures, by examination, the performance of individuals against the task description of the rate and the performance of schools against the teaching tasks they have been assigned. Individuals are tested semi-annually using a "System Achievement Test". End of course examinations or "Course Achievement Tests" are provided for the schools. The tests are generated by computer from a bank of some 17,000 test items. Scoring is also done by the computer and statistical analysis done on tests and test items. This test and test item analysis forms the basis of the evaluations performed by PTEP. In conducting evaluations PTEP also draws upon computer stored personnel training information. Records are available on the approximately 4000 FBM Weapons System personnel. (Figure 2)

Evaluations have been conducted on several of the aspects of replacement and advanced training. Performance in an area of interest on a System Achievement Test has been the base line in each case and a delta across a course or time interval has then been compared with a control group figure. By the process we have measured:

1. Performance of courses against established standards.

2. Performance of individuals with varying degrees of preschool knowledge in a course.

3. The amount of advanced training material addressed

by replacement training.

4. Practicality of replacement training for retraining senior petty officers on updated equipment.

5. The ability of a course to support maintenance requirements.

Each of these areas are areas where evaluation findings can be applied in the saving of training time or dollars or in the better training support of shipboard equipment. We have learned things of specific applicability in each of the above cases but more importantly, from the testing point of view, we learned a great deal about the limitations of small population sampling.

Our System Achievement Tests run from 180 to 360 multiple choice questions depending on the Navy Enlisted Classification (NEC) to which they apply. Each test is subdivided into areas of equipment responsibility. In our earlier tests we worked with a minimum area size of 11 test items. The tests are designed for average score of 50% correct. The advanced courses in Fleet Ballistic Missile Weapons training normally have 5 to 6 men per class, run for one or two weeks and convene on an 8 week cycle.

It was early noted in our investigation that because of the small class/sample sizes that our evaluations were influenced by considerations that on larger samples could be neglected. For example:

1. Because of the long time required to accumulate a usable sample, continuing relevance of the conclusions had to be considered.

2. The small number of questions in an area did not support other than general conclusions.

3. Extremes of performance by one or two individuals had a large impact on averages.

4. Selecting control groups on a pairing basis, using only a single performance criteria, could easily yield an control group which was not truly representative of the population at large.

It was apparent that some largely subjective performance standards had to be applied if results were to be published which could be defended in good conscience. Some simple procedures were adopted:

1. Tests were reexamined item by item for test item performance and relevance.

2. Personal training data was examined on personnel with extremes of performance in an area to see if the performance was consistent with their training record.

3. For very small groups personnel data was examined for the entire group.

4. Control groups which were greater in size than the study group were used to better define the "untreated" population.

5. Raw scores were examined in addition to percentages or standardized scores to give a better feel for the size of a change.

To help in future evaluations, minimum area size in System Achievement Tests was increased from 11 to 18 questions and an effort made to use at least 30 questions per area.

The examination of the objective data in the light of the more subjective information in some cases left the evaluation result in the same sense as the test result and in other cases produced diametric changes in conclusions. The combination, however, produced reliable evaluation results which can be put forward and acted on with some confidence.

Sampling tasks responsibility with a limited number of questions has proved to be an excellent way of indicating problem areas, monitoring for consistent performance and showing trends. For problem definition so that corrective action can be taken test results must be supplemented with personnel and technical data.

The Interface of Testing and Training: Air Force Experience 1972 - 1974

by Lyle D. Kaapke, Colonel, USAF

Introduction

The theme of the 1974 MTA Conference is particularly appropriate to what has been happening in Air Force testing and training programs in the last few years and gives us an excellent opportunity to review where we are and where we are going with these programs.

In 1969, the test development program in support of the Weighted Airman Promotion System (WAPS) was moved from the Personnel Research Division of AFHRL, Air Force Systems Command, and reassigned to Air Training Command where it was grouped with the occupational survey program as Detachment 17 of HQ ATC. The transition of operational test development from a research command to the training command implicitly recognized the common bond which exists between testing and training in the Air Force.

The USAF Occupational Measurement Center

Before I address the past experience we have had in facilitating and strengthening the bond between testing and training, I would like to present a short overview of our present organization, its missions, and concept of operation.

Mission

Our dual mission as depicted on our Center Emblem is to measure the job and measure the man.

We are specifically charged with the conduct and completion of occupational surveys and the development of personnel tests largely in support of the Weighted Airman Promotion System.

Organization

Our organization is functionally configured to accomplish its dual mission. Our personnel represent an interesting mix of civilian and military -- professional level and support personnel. We are authorized 127 people -- 38 officers, 26 enlisted, and 63 civilian. The 38 officers and approximately 30 civilians are largely behavioral scientists, possessing Master's or higher academic degrees.

The test services branch is responsible for minor test projects and the development branch major test projects, which I will explain shortly.

Occupational Survey Program

I would like to take a moment to describe the highlights of our occupational survey program. Surveys of 51 enlisted career fields are accomplished each year to determine the specific tasks performed in a specialty and the relative amount of time spent performing them. Dr. Driskill and others will elaborate in more detail so I'll just give you a general idea how all this fits together.

Results of the Survey Program

Reports of results after computer analysis provide recommendations concerning technical training, classification, and career field structure.

Savings or averted training costs since the program's inception in 1967 amount to some 12 million dollars and 160,000 mandays of training time.

The Test Development Program

Let's turn now to the mission of test development in support of WAPS.

I would like to take a moment to describe our concept of test construction before we consider the composition of specific promotion instruments themselves.

The acceptance and approval of the WAPS promotion program results in large part from the role airmen play in test development activities.

Team Rationale

Highly qualified, knowledgeable NCO subject-matter specialists, in the grade of Master Sergeant or above perform TDY to this Center. They work with civilian or military test development psychologists and each adds his or her own special expertise in terms of subject matter or test development methodology to produce an up-to-date, relevant, and discerning measure of required knowledge.

This interaction of practical, relevant, on-the-job knowledge and expertise in test development methodology results in the construction of superior promotion instruments.

The Specialty Knowledge Test

One of these instruments is the Specialty Knowledge Test. These tests, presently covering approximately 200 career fields and their shreds, are revised once or more each year. SKTs are prepared according to skill level knowledge requirements.

At least every other year, or more often if needed, a thorough overhaul of a career field's tests is completed using five or more subject-matter specialists.

Substantially larger teams are needed if the career field contains many shreds in addition to the basic specialty. Major test construction projects are accomplished by the test development branch.

In the off year, a test is revised by two subject-matter specialists and a test development psychologist who replace items made obsolete by advances in technology, changes in equipment and function, or improvement in methods and procedures. The minor project as well as SMS procurement and final test preparation are accomplished by the test services branch

These two test development approaches -- the major or minor project -- permit us to meet Air Force goals for the annual revision of all tests.

The Promotion Fitness Examination

Another principal promotion instrument is the Promotion Fitness Exam which contains items covering the knowledge areas shown. It is designed to be given at appropriate grade levels.

The PFE and the Supervisory Exam are totally reworked once a year. Team members are all senior or chief master sergeants nominated by their respective commands and selected at HQ USAF. Generally, there are 12 or more active duty members and a representative from each, the Air Guard and Air Reserve.

Test Development Program Results

With the assistance of over 800 subject-matter specialists working in concert with our permanently assigned personnel using the techniques and concepts I have described, we were able to produce over 1,000 tests in CYs 72 and 73, and meet the Air Force's goal for annual revision of all promotion test instruments. Then Air Force Chief of Staff, General Ryan, remarked that our organization had more to do with the confidence and well-being of the entire enlisted Air Force population than any other unit he was aware of.

The Training-Testing Bond

Now that we have a common understanding of how the organization functions, I would like to specifically discuss the efforts we have taken during my tenure with the organization to provide a smoother interface between testing and training. Going back a bit, on 1 January 1972, Detachment 17 became the 3700 Occupational Measurement Squadron under the Air Force Military Training Center. During 1972 and 1973, the unit worked closely with the technical training centers and HQ ATC to program and develop occupational surveys as key elements in the establishment of training requirements and career field structures. At the same time, participation of training personnel from the technical training centers in test development projects was maximized by requiring the CDC writer to participate in each SKT project.

Training-Testing Compatibility

We could epitomize the last couple of years as ones in which we have strived to achieve compatibility. We put training and testing on a common basis by using the same documents as the basic source of training and testing developments. There is a level of abstraction concerning occupations which runs from the most general found in the 39-1 job classification manual to the specific standards of proficiency or knowledge represented in a career ladder's STS to what tasks are actually performed in the field as found in the occupational survey report. These documents and information are used to develop technical training course curricula as well as to develop test outlines of the areas to be covered by an SKT. So the starting point is largely the same whether one is developing a training course or a knowledge test of some specialty.

The Career Development Course has long been used to support skill upgrading or a two Channel OJT Approach. One channel is developing proficiency in a specific job assignment, the other acquiring the broad knowledge of the entire specialty found in the CDC. We create greater compatibility by also using the CDC where possible as the sole study reference source for an SKT. The concept underlying the attempt to provide compatibility between the two systems is that benefits will derive from perceived system credibility and increased motivation on the part of the enlisted population.

An individual studies the knowledge areas outlined on the STS at Tech School and later in OJT. He studies his CDC in order to increase his skill level. Finally, he reaches a grade where he must take an SKT to be eligible for promotion. The sole reference he must study is again the most up-to-date version of his CDC. When the individual takes his SKT, he should have a feeling of continuity and great understanding of the material being tested -- much of it is common to his Tech School training, his skill upgrading exam, and his subsequent study. The extent to which the individual has an "Aha" experience -- it suddenly dawns on him that the whole system is in gear, makes sense, and has continuity -- is the extent to which we have achieved desired compatibility between the subsystems. Thus, each subsystem supports and stimulates activity in the other and the interface between them provides great credibility for both.

Payoffs for Training and Testing

The payoff associated with a close interface between training and testing has been substantial. As previously stated, the CDC writer is a member of all major test revision teams. We find that his immediate interaction with and exposure to the attitudes of fellow team members regarding the currency and comprehensiveness of the CDC results in a more adequate basic study reference. This directly helps the skill upgrading program by infusing the latest operationally oriented information into the program through the CDC. The CDC writer often brings in an accepted manuscript or authorized revision to be used in test development -- this insures that the publication of a new CDC precedes the test based on that material only by the period necessary for study. Thus, current, up-to-date and reliable information is used for both tests and study material for skill upgrading.

We also studied the benefits to be derived from the preparation of SKTs at the appropriate Tech Center under the direction of a project officer trained at our unit but assigned to the Center. The cross fertilization of ideas from training and the field resulted in appreciable benefits for both training and testing and the project is continuing on a selective basis. We found that test development at the individual Tech Centers could best be performed in areas with special problems such as nonavailability of subject-matter specialists, no previous SKTs in existence, or where there are substantial prospective revisions to the STS or Specialty description. We are gratified that this additional avenue is available and that the Tech Centers are providing such eager support for these special test development efforts. The Tech Training Center profits from the evaluation of training courses provided by team members who possess outstanding knowledge of the specialty and the current situation which exists in the field vis-a-vis that specialty.

On the other hand, SKT projects in-house provide a payoff to training in the form of detailed objective evaluations of specialty descriptions and training documents as well as the Career Development Course itself. As one team recently put it, we had a hundred and forty years of career field experience represented. Seldom are we able to assemble such a wealth of highly qualified people to take an in-depth look at specialty training documents. We presently attach all study reference requirements to the STS so that all the training and testing requirements are available to every enlisted man as part of his personal OJT folder. Finally, the specialty experience represented on the team can be drawn upon within our own organization by occupational analysts and inventory developers. This saves a good chunk of TDY costs since we literally serve a double purpose in having the team come to the military developer rather than vice versa. We aren't able to work the program requirements in test development and survey perfectly, but have found the programs support each other to a high degree.

Changes in Mission and Organization

We have seen where we have been and how the organization has developed up to a point. Now, it would be appropriate to consider where we are going from here. A number of significant events occurred on 1 July of this year. Our unit was redesignated the USAF Occupational Measurement Center. We also reorganized to provide improved managerial lines of responsibility, received new authorizations of fourteen positions and initiated a program to develop tests for the Air Guard and Reserve in support of a Weighted Airman Screening System. The test program is now an integral part of the total force concept -- a merging of all Air Force resources and responsibilities or missions. Our redesignation has improved lines of communication and the support of the major commands in responding to SMS levies and survey requirements. We are currently involved in the field survey portion of a study of the structure of the Communications-Electronics Officer career field. This is the first officer survey we have attempted on an operational basis.

Preliminary results would seem to suggest that the officer area will be as fertile as that of the enlisted career ladders. Perhaps Dr. Driskill will tell you more about that.

We have some equipment features that have further streamlined the preparation of tests -- the storage and preparation of test copy on magnetic tape compatible to composer equipment that will prepare dual column formating and reduce paper and printing requirements by a third.

This has been a very fast look at our Center -- where we have been, where we are going, and how we hope to get there. I've had to gloss over many important areas but hope I have given you a view of a viable, moving, progressive test development and survey facility that dovetails with and is fully compatible to an extensive training system -- one that has credibility and meaning for every Air Force enlisted member. May I in the time remaining entertain your questions.

SECTION XI

COMPUTER ASSISTED TRAINING AND TESTING

CONTRIBUTIONS TO OBJECTIVE MEASUREMENT AND EVALUATION OF TRAINEE COMPETENCY

by

William J. Moonan, Ph.D.

Director, Statistical Programs
Navy Personnel Research and Development Center
San Diego, California 92152

A Presentation Made to

The 1974 Military Testing Association Conference
Oklahoma City, Oklahoma

21-25 October 1974

TABLE OF CONTENTS

Page

1. Introduction 1

2. Background 1

3. The Binary Measurement Model 5

4. Model Objectivity 6

   Concept No. 1. 6

   Implication No. 1. 6

   Concept No. 2 7

   Implication No. 2 7

5. The Frequency Ratio Method (FRM) 11

6. A Numerical Example 14

   Phase I 14

   Phase II 15

7. Estimation of Competency 20

8. The COMET System 22

   Description of Sub-System Functions 23

9. Expected Implementation Problems 26

10. Summary 27

11. References 27

LIST OF EXHIBITS

Exhibit		Page
1	Types of Measurement Models Developed by Rasch.	18
2	Item Easiness Estimation Procedures	18
3	Stepwise Procedure for FRM (Phase I), Item Easiness Estimation Computations	19
4	Stepwise Procedure for FRM (Phase II), Goodness of Fit Computation	19
5	Competency Estimates for an Examination of I=K=7 ETST Items	21

1. Introduction

The purpose of this paper is to lay a basis for and discuss the components of a system, called COMET, designed to objectively measure and evaluate the competency of trainees in military training enterprises. COMET is an acronym for "Computerized Objective Measurement and Evaluation of Trainees."

These goals will be accomplished by:

a) describing what the author means by the terms "measurement," "objectivity," and "evaluation," by indicating systems antecedent to COMET, and describing the objective Binary Measurement Model (BMM) used in the COMET system,

b) deriving and illustrating a new procedure for estimating one of the parameters and fit of the BMM model, called the Frequency Ratio Method (FRM), and

c) specifying the component subsystems, and their functions, of COMET as well as its goal, objectives and practical problems associated with its implementation.

A detailed numerical example is provided to facilitate understanding of COMET's mathematical and statistical aspects.

2. Background

By the term "measurement" the author refers to the process by which numbers are assigned to a property of an entity. In training enterprises, the entities are personnel or trainees whereas the property being measured we refer to as "competency" in the training field.

By the term "evaluation" the author refers to the process by which measurements or their functions are used to characterize features or aspects of the evaluated "situation" or the entities included therein.

For example, we may evaluate a trainee (entity) during training (situation) by the process of taking the difference (function) of his competency (measurement) as determined at two different stages of training. Notice that no other value judgment is placed on the evaluation result in this definition. This was omitted because the value applied to an evaluation result is not unique. It is different depending upon the decision maker and the decision problem wherein the evaluations are to be utilized for decision purposes. If the results are not to be utilized for decision purposes, they need no value judgment component.

By the term "objective" the author means different things depending upon what is referred to. There are three different senses in which the term is used in this paper. They are

a) Training program objectivity,

b) Measurement Instrument objectivity, and

c) Measurement model parameter objectivity.

Training Program Objectivity refers to the process whereby training programs are objectively established by means of expressly stated goals, functions and objectives and are designed and managed by a carefully thought out, systems approach such as that developed by Rundquist (1967).

Measurement Instrument Objectivity refers to process whereby a measurement instrument (test) is designed to be free of ambiguity and subjectivity in its administration or in responding to or scoring its items.

Measurement Model Parameter Objectivity refers to two properties of a measurement model, viz.

a) the comparison of any two testees may be carried out in such a way that no other measurement parameters are involved than those of the two testees, and

b) any two items may be compared independently of all others parameters (including those of the testees) than those of the two items.

We emphasize the desirability of requiring an evaluation system, automated or not, to possess the three aspects of objectivity, otherwise it is likely not to meet its goal and objectives as well as not providing useful measurements and evaluations.

As far as is known, there is only one class of measurement model (those developed by Rasch (1960) which possess the properties of measurement model objectivity. That these properties actually exist for some of the Rasch models, has been shown by Rasch himself (1969), Wright (1969), Choppin (1968), and Schmidt (1969). Later in this paper we will provide yet another demonstration. A most startling and readable account of the Rasch BMM's objective features has been provided by Wright (1968). My own account, Moonan (1972), may be of use and conveniently available to you.

We are aware of at least two antecedent, automated trainee evaluation systems. There are probably others, but our attention has not been directed to them. The designers deserve our applause for their pioneering work and we can only wish that the COMET system we propose could be implemented as effectively and successfully as each of these were. The first system, I refer to as "GLAKES," was developed by CDR W. H. Wheeler while stationed at the Navy Electronic Technician's School at the Naval Training Center, Great Lakes, Illinois. My documentation, Wheeler (1966) is meager, but I did see it in operation. I was particularly impressed by its test reproduction subsystem wherein examinations were quickly and conveniently assembled from the item files. Its evaluation subsystem was also excellently designed.

The second system was developed for the Marine Corps Communications Electronics School at the Marine Corps Recruit Depot, San Diego by Dr. Richard S. Hatch, (Decisions Systems Associates, Inc., Rockville, MD). This system is known by the acronym ATAC; Automatic Testing and Attrition Control and, as its name implies, contained a subsystem for controlling how many, which, and when trainees should be attrited from the training program. This system had other excellent features as well. Most particular were its reporting subsystem and the evaluation subsystem.

Both systems were ahead of their time, the BMM was not well known nor computationally effective at the time of their design. Their development shows that only "bare bones" design features can, at first, be implemented but must ultimately be augmented by features which fulfill local requirements, overcome constraints and meet local desires. It is believed that the basic COMET system, described herein, supplies the essential "bare bones" for many implementation situations.

It should be mentioned here, most fittingly, that my own particular interest in this research area was substantially motivated by seeing CLAKES in operation, working with Dr. Hatch on ATAC and being encouraged by CAPT O. L. Dawson, USCG, to develop an effective COMET-like system for the U.S. Coast Guard Training Facility in New York.

My personal documentation on ATA is limited, Blakely (1969), and Hatch (1967) but, perhaps Dr. Hatch could supply additional information if it is needed.

3. The Binary Measurement Model

Rasch (1960) developed, essentially, three sub-classes of a general class of objective measurement models. I refer to these models as CD, AD and SP models as described in Exhibit 1. Actually the AD model is sufficiently general to encompass the CD model which in turn can be shown to mathematically encompass SP models. The BMM is the AD model and is the one discussed previously (Moonan, 1972). The mathematical statement of the model is expressed in probability terms. The probability of a response, $X(i,j)$ being correct, $X_{(i,j)} = 1$, when made to the ith item, with easiness, $\varepsilon(i)$, by the nth trainee who has competency $\gamma(n)$ is given by the expression, where $j=n$,

$$(1) \qquad P[X_{(i,j)} = 1 \mid \varepsilon(i),\ \gamma(j)] = \frac{\varepsilon(i) \cdot \gamma(n)}{1 + \varepsilon(i) \cdot \gamma(n)} = P_1 \quad .$$

Similarly, the probability of an incorrect response is given by

$$(2) \qquad P[X_{(i,j)} = 0 \mid \varepsilon(i),\ \gamma(n)] = \frac{1}{1 + \varepsilon(i) \cdot \gamma(n)} = P_0 = 1 - P_1 \quad .$$

The model is called "Binary" because responses to items are only allowed to be binary or dichotomous. Responses to Affective Domain items are usually n-ary so that the Rasch polychotomous model is more appropriate for measurement under such conditions. In this case, when $n=2$, the CD:BMM model is equivalent to the AD sub-model which we will not discuss further at this time.

The model produces "measurements" because its use permits the assignment of numbers to the "competency" of trainees. It also "measures" the items since the easiness, $\varepsilon(i)$, of each item may be numerically estimated.

The scale of competency has a zero point and an estimable unit of measurement. Thus, the model, with its objective properties provides a suitable measurement basis for the COMET system.

The BMM has two parameters, $\gamma(j)$ and $\varepsilon(i)$. The first we call "competency" rather than the more usual "ability." We prefer the former term for training situations and have avoided the confusion sometimes resulting from the word "ability." The easiness parameter estimated indicates how "easy" or "difficult" an item's correct response is to discern for a testee.

4. Model Objectivity

As was pointed out in earlier sections, Objectivity plays an important role in the COMET system. In this section we shall examine the model objectivity of the BMM by considering the dualistic concepts and implications of measurement model objectivity. These duals are listed below:

Concept No. 1. The item parameters, $\varepsilon(i)$, can be estimated independently and without knowledge of the competency parameters, $\gamma(n)$, of the persons responding to the items.

Implication No. 1. The easiness scale can be calibrated and an easiness can be estimated independently of the persons, and their competencies, responding to the items. This implication is very important for the COMET system for it means we can collect item responses data from whomever it is convenient to do so. Items can be administered in small sets of items to persons of quite divergent competencies. Groups of similar persons are not required to respond to the same items.

Concept No. 2. The competency parameters, $\gamma(n)$, can be estimated independently and without knowledge of the item parameters, $\varepsilon(i)$.

Implication No. 2. The competency scale can be calibrated and a competency can be measured independently of the items used for measurement. Simply stated this means that, if desired, individualized tests for each trainee could be used, but in doing so, each trainee is measured on the same competency scale.

We shall mathematically prove the two concept statements for the BMM below:

According to the model we have the single item probabilities

$$P[X(i,n) = 1] = \frac{\varepsilon(i) \cdot \gamma(n)}{1 + \varepsilon(i) \cdot \gamma(n)} \tag{3}$$

, and

$$P[X(i,n) = 0] = \frac{1}{1 + \varepsilon(i) \cdot \gamma(n)} \tag{4}$$

.

Consider next the probabilities associated with the four possible events which can occur if the same person, n, responds to two items, i and j. To this end let,

(5) $D(i) = 1 + \varepsilon(i)\,\gamma(n)$ and $D(j) = 1 + \varepsilon(j) \cdot \gamma(n)$ be the denominator terms of our probability expressions, then (assuming response independence)

$$P[X(i,n) = 0,\ X(j,n) = 0] = \frac{1}{D(i) \cdot D(j)} \tag{6}$$

,

$$P[X(i,n) = 1,\ X(j,n) = 0] = \frac{\varepsilon(i) \cdot \gamma(n)}{D(i) \cdot D(j)} \tag{7}$$

,

$$P[X(i,n) = 0,\ X(j,n) = 1] = \frac{\varepsilon(j) \cdot \gamma(n)}{D(i) \cdot D(j)} \tag{8}$$

, and lastly

$$(9)\quad P[X(i,n) = 1,\ X(j,n) = 1] = \frac{\varepsilon(i) \cdot \gamma(n) \cdot \varepsilon(j) \cdot \gamma(n)}{D(i) \cdot D(j)} .$$

The conditional probability that the nth person gets the ith item correct, given that he scored a total of 1 on both item i and j is given by the equation ratio (11) = (8)/(10) where (10) is given by the equation sum (10) = (7) + (8). Thus

$$(10)\quad P[X(i,n) + X(j,n) = 1] = \frac{\{\varepsilon(i) + \varepsilon(j)\} \cdot \gamma(n)}{D(i) \cdot D(j)} ,\ \text{and}$$

$$(11)\quad \frac{(8)}{(10)} = \frac{\varepsilon(i)}{\varepsilon(i) + \varepsilon(j)}$$ in which we note the $\gamma(n)$ cancels,

and which indicates the truth of Concept No. 1, for suppose we have collcected responses made by a set of persons to items i and j, and suppose that

(12) $f(i,j)$ of them got item i correct and item j incorrect; and

(13) $f(j,k)$ of them got item j correct and item i incorrect, we note by

(14) $g(i,j) = f(i,j) + f(j,i)$ the number of these persons who receive a total score of 1 on both of the items.

We note, additionally, that

$$(15)\quad \frac{f(i,j)}{g(i,j)} \text{ estimates } \frac{\varepsilon(i)}{\varepsilon(i) + \varepsilon(j)} ,$$

$$(16)\quad f(j,i) \text{ estimates } \frac{\varepsilon(j)}{\varepsilon(i) + \varepsilon(j)} ,\ \text{and}$$

$$(17)\quad \frac{f(i,j)}{f(j,i)} \text{ estimates } \frac{\varepsilon(i)}{\varepsilon(j)}$$, further substantiating Concept No. 1.

Eqn (17) was derived by Rasch (1970), Choppin (1968), and others, but none, to my knowledge, have used this fact to develop an $\varepsilon(i)$ estimation method, based on the frequency ratios. This process known as the Frequency Ratio Method (FRM) represents the author's modest "contribution" to objective measurement and forms a computational basis for the COMET system. Exhibit 2 shows some available item easiness computational methods.

To continue our derivations consider an item, i, responded to by two persons with different competencies, $\gamma(m)$ and $\gamma(n)$. Again there are 4 outcomes and their probabilities are listed below:

(18) Let $D(m) = 1 + \varepsilon(i) \cdot \gamma(m)$ and $D(n) = 1 + \varepsilon(i) \cdot \gamma(n)$ be the denominator terms in the probability expressions

$$(19) \quad P[X(m,i) = 0,\ X(n,i) = 0] = \frac{1}{D(m) \cdot D(n)}$$

$$(20) \quad P[X(m,i) = 1,\ X(n,i) = 0] = \frac{\varepsilon(i) \cdot \gamma(m)}{D(m) \cdot D(n)}$$

$$(21) \quad P[X(m,i) = 0,\ X(n,i) = 1] = \frac{\varepsilon(i) \cdot \gamma(n)}{D(m) \cdot D(n)}$$

$$(22) \quad P[X(m,i) = 1,\ X(n,i) = 1] = \frac{\varepsilon^2(i) \cdot \gamma(m) \cdot \gamma(n)}{D(m) \cdot D(n)}$$

The conditional probability that the ith item was correctly responded to by the mth person given that the total score made by the mth and nth persons to that item was equal to 1 is given by the equation ratio (24) = (20)/(23) where (23) is given by the equation sum (23)=(20)+(21), thus

$$(23) \quad P[X(m,i) + X(n,i) = 1] = \frac{\varepsilon(i) \cdot \{\gamma(m) + \gamma(n)\}}{D(m) \cdot D(n)} \text{ , and}$$

$$(24) \qquad \frac{(20)}{(23)} = \frac{\gamma(m)}{\gamma(m) + \gamma(n)}$$

which indicates the truth of Concept No. 2, for suppose we have collected responses made to a set of items by persons m and n and suppose

(25) $F(m,n)$ of the items were correctly responded to by person m and were incorrectly responded to by person n, and

(26) $F(n,m)$ of the items were correctly responded to by person n and were incorrectly responded to by person m. We also let

(27) $G(m,n) = F(m,n) + F(n,m)$ be the number of items which were responded to correctly by persons m or n.

We can see immediately that

(28) $\frac{F(m,n)}{G(m,n)}$ estimates $\frac{\gamma(m)}{\gamma(m) + \gamma(n)}$, and

(29) $\frac{F(n,m)}{G(m,n)}$ estimates $\frac{\gamma(n)}{\gamma(m) + \gamma(n)}$, whereas

(30) $\frac{F(m,n)}{F(n,m)}$ estimates $\frac{\gamma(m)}{\gamma(n)}$, further substantiating Concept No. 2.

Rasch (1970) indicates that (30) cannot be practically used to estimate $\gamma(n)$ since the number of items used in an assessment program is usually much smaller than the number of testees. Consequently we will use another approach to estimate $\gamma(n)$ in the FRM or the measurement model analysis sub-system of COMET, although, it seems clear, that in some assessment situations, the number of items is sufficient if the number of $\gamma(n)$ to be estimated corresponds to the number of possible raw scores.

5. The Frequency Ratio Method (FRM)

The FRM is a mathematical-computational procedure designed to provide a simple estimation and goodness of fit test for the BMM parameters and responses. It is intended to be an analytical component of the COMET system. To illustrate FRM we assume we have collected from trainee-persons their responses to the available items of an item bank constructed from the learning (training) objectives of the training course and assumed to indicate training competency if the items are correctly responded to by the trainees. Our available data can be portrayed as the frequencies of equations (12) and (13) and arrayed in a square matrix R(i,j) whose main diagonal is left blank.

After portraying the mathematics and statistics of the FRM we will formulate the computations in two stepwise phases.

Phase I: Item easiness estimation computations.

Phase II: Goodness of Fit computations.

In R(i,j) we form ratios of symmetrical elements and form the square matrix S(i,j). Thus

$$(31) \quad R(i,j) = \frac{f(i,j)}{f(j,i)} \quad \text{for} \quad i>j,\ i=j=1,\ \ldots,\ I,$$

$$(32) \quad R(j,i) = \frac{f(j,i)}{f(i,j)} \quad \text{for} \quad j>i,\ j=i=1,\ \ldots,\ I, \text{ and}$$

$$(33) \quad R(i,j) = \text{blank for } i=1,\ \ldots,\ I.$$

We know from (17) that (31) estimates $\varepsilon(i)/\varepsilon(j)$ and (32) estimates $\varepsilon(j)/\varepsilon(i)$ so that the natural logarithm of (31) is

(34) $\ln S(i,j) = \ln \varepsilon(i) - \ln \varepsilon(j)$ whereas for (32) we have

(35) $\ln S(j,i) = \ln \varepsilon(j) - \ln \varepsilon(i) = -\ln S(i,j)$

If we sum the j^{th} row of $\ln S(i,j)$ we get, after adding and subtracting $\varepsilon(j)$,

(36) $T(j) = \sum_{i=1}^{I} \ln \varepsilon(i) - I \ln \varepsilon(j)$ and let

(37) $\sum_{i=1}^{I} \ln \varepsilon(i) = 0$ for definiteness; and

(38) $\ln \varepsilon(j) = \frac{-T(j)}{I} = t(j)$, and $j = 1, \ldots, I$, so that

(39) $\hat{\varepsilon}(j) = e^{t(j)}$.

A stepwise computational procedure is indicated in Exhibit 3. This simple estimation procedure requires only the development and maintenence of a square matrix of frequencies whose elements are the numbers of those persons tested on some of the items and who responded correctly to one, and only one, of each pair of items presented to them. A list of item easiness estimation procedures are shown in Exhibit 2.

Eqn (39) is the culmination of the item easiness estimation phase of the FRM. We now proceed to develop the second phase which is intended to provide a goodness of fit test of the response data to the BMM.

With (39) we are provided with estimates, $\hat{\varepsilon}(i)$, of the easiness parameters, $\varepsilon(i)$. We intend to use the χ^2 distribution (Fisher (1948), as the basis of our goodness of fit test. The hypothesis tested is that the items used for assessment of competency evoke responses in accordance with the BMM. The chi-square form we use is

$$\chi^2(i,j) = [f(i,j) - e(i,j)]^2/e(ij) \quad . \tag{40}$$

using (12) and (13) for the f(i,j) observed frequencies and substantiating (39) in (15) we define the expected frequencies as

$$e(i,j) = g(i,j) \cdot \frac{\hat{\varepsilon}(i)}{\hat{\varepsilon}(i) + \hat{\varepsilon}(j)} \text{ , } i>j\text{, and} \tag{41}$$

$$e(j,i) = g(i,j) \cdot \frac{\hat{\varepsilon}(j)}{\hat{\varepsilon}(i) + \hat{\varepsilon}(j)} \text{ , } j>i \quad . \tag{42}$$

We can compute χ^2 contributions to the upper and lower parts with the following formalae

$$\chi^2(i,j) = [f(i,j) - e(i,j)]^2/e(i,j)\text{, } i>j\text{, and} \tag{43}$$

$$\chi^2(j,i) = [f(j,i) - e(j,i)]^2/e(j,i)\text{, } j>i \quad . \tag{44}$$

We next add these values symmetrically into a single triangular array by the formula

$$C(i,j) = \chi^2(i,j) + \chi^2(j,i) \quad , \tag{45}$$

considered as an array with a total of I(I-2) degrees of freedom, so that were we to compute a $\chi^2(k)$ for each item from this table, it would have I-2 degrees of freedom. We compute $\chi^2(k)$ from C(i,j) by summing elements in an "L-shaped" fashion according to the following formula

$$\chi^2(k) = \sum_{j=1}^{k-1} C(i,k) + \sum_{i=k+1}^{I} C(k,j)\text{, } k=1, \ldots, I \quad . \tag{46}$$

If any $\chi^2(k)$ exceeds the tabled χ^2 for I-2 degrees of freedom at the selected α level, the hypothesis that the responses to the item fit the BMM is rejected and the item should be removed from the COMET system.

Exhibit 4 shows the stepwise computations for these FRM (phase II) goodness of fit calculations.

6. A Numerical Example (Phase 1)

Step

0 This illustration is intended to show the numerical aspects resulting from applying the FRM to a set of response data. The data originated from the responses of Naval recruits to the first I=7 items of the Electronic Technician Selection Test (ETST). Originally 9 items were used but the 8th and 9th item responses were found not to fit the BMM model so they were omitted from the analysis. All recruits responded to the same 7 items so the data does not result from the "catch as catch can" response sampling procedure mentioned earlier. This difference causes no problem in our analysis.

1 The S(i,j) matrix of frequencies

	ITEM INCORRECT						
ITEM CORRECT	1	2	3	4	5	6	7
1	-	220	210	97	203	202	269
2	118	-	157	87	163	148	198
3	111	160	-	62	164	147	205
4	183	275	248	-	277	247	330
5	85	147	145	73	-	130	202
6	164	212	208	123	210	-	245
7	113	144	148	88	164	127	-

2 Compute Frequency Ratios R(i,j)

	1	2	3	4	5	6	7
1	-	1.864	1.892	.530	2.388	1.232	2.381
2	.536	-	.981	.316	1.109	.698	1.375
3	.529	1.019	-	.250	1.131	.707	1.385
4	.887	3.161	4.000	-	3.795	2.008	3.750
5	.419	.902	.884	.264	-	.619	1.232
6	.812	1.432	1.415	.498	1.615	-	1.929
7	.420	.727	.722	.267	,812	.518	-

Step

3 T(i,j) = ln R(i,j)

	1	2	3	4	5	6	7
1	-	.622	.638	-.635	.870	.208	.867
2	-.622	-	-.019	-1.151	.103	-.359	.318
3	-.638	.019	-	1.386	.123	-.347	.326
4	.635	1.151	1.386	-	1.331	.697	1.321
5	-.870	-.103	-.123	-1.331	-	-.480	.208
6	-.208	.359	.347	-.697	.480	-	.657
7	-.867	-.318	-.326	-1.321	-.208	-.657	-

4,5,6 Compute T(j), t(i) and $\hat{\varepsilon}(j)$

	T(j)	t(j)	$\hat{\varepsilon}(j)$
1	2.570	-.367	.693
2	-1.730	.247	1.280
3	-1.903	.272	1.313
4	6.521	-.932	.394
5	-2.699	.386	1.471
6	.438	-.134	.874
7	-3.697	.528	1.696

This completes Phase 1.

Phase 2 - Goodness of Fit Test

Step

0 the $\hat{\varepsilon}(i)$ values are given above in Step 6 of Phase 1

1 the S(ij) matrix, containing f(i,j) and f(j,i), is given in Step 1 of Phase (1). The g(i,j) matrix is given below

	1	2	3	4	5	6	7
1	--	338	321	280	288	366	782
2		--	317	362	310	360	342
3			—	310	309	355	353
4				—	350	370	418
5	(Symmetrical)				--	340	366
6						--	372
7							--

Step

2 Expected Frequencies

	1	2	3	4	5	6	7
1		219.29	210.09	101.50	195.78	204.15	271.18
2	118.71	--	160.53	85.21	165.76	146.09	194.91
3	110.91	156.47	--	71.55	163.28	141.86	198.95
4	178.50	276.79	238.45	--	276.05	255.04	339.21
5	92.23	143.31	145.72	73.94	--	126.72	195.99
6	161.85	213.91	213.14	114.95	213.28	--	245.48
7	110.82	147.09	154.05	78.79	170.01	126.52	--

3 χ^2 Contributions

	1	2	3	4	5	6	7
1	--	.002	.000	.200	.266	.023	.018
2	.004	--	.078	.038	.046	.025	.049
3	.000	.080	--	1.275	.003	.186	.184
4	.113	.112	.382	--	.003	.253	.250
5	.565	.053	.004	.012	--	.085	.184
6	.029	.017	.124	.562	.050	--	.000
7	.043	.065	.238	1.077	.212	.002	--

4 Accumulated χ^2 values

	1	2	3	4	5	6	7
1	--	.006	.000	.313	.831	.052	.061
2		--	.158	.050	.099	.042	.114
3			--	1.657	.007	.031	.422
4				--	.015	.815	1.327
5					--	.002	1.077
6						--	2.322
7							--

5 $\alpha = .05$, $df = I-2 = 5$ $\chi^2(.95, 5) = 11.070 = \chi^2_o$

Step

6 Compute $\chi^2(k)$

k	$\chi^2(k)$
1	1.263
2	.469
3	2.275
4	4.177
5	2.031
6	2.019
7	5.323

7 No $\chi^2(k)$ exceeds or equals χ_0^2, therefore H_0 is accepted for all 7 items.

8 As no items are removed, it is not necessary to reestimate the $\hat{c}(k)$.

This completes Phase 2.

EXHIBIT 1

Types of Measurement Models Developed by Rasch (1960)

Type	Name	Application Fields
CD	Cognitive Domain	Measurement of Intelligence, Aptitudes, Competencies and Abilities
AD	Affective Domain	Measurement of Attitudes, Values Motivation and Interests
SP	Stochastic Process	Measurement involving stochastic process observations such as reading errors

EXHIBIT 2

Item Easiness Estimation Procedures

Method Name	Originator
Least Squares	Bramble (1969)
Graphic	Rasch (1960)
Maximum Likelihood	Wright and Panchapakesan (1969)
Logarithmic	Blommers (1965)
Frequency Ratio	Moonan (1974)

EXHIBIT 3

Stepwise Procedure for FRM (Phase I)
Item Easiness Estimation Computations

Step

0 Collect responses to the items in the item competency bank.

1 Organize the responses into a square matrix S(i,j) whose elements are defined by (12), (13) and (14).

2 Compute the frequency ratios as indicated by (31) and (32). Call the matrix of ratios R(i,j).

3 Take the natural logarithms of the R(i,j) elements; called T(i,j).

4 Sum each row of the matrix of logarithms and store in a vector T(j).

5 Compute $t(j) = -T(j)/(I-1)$

6 Determine the easiness estimates by the relation $\hat{\varepsilon}(j) = e^{t(j)}$.

EXHIBIT 4

Stepwise Procedure for FRM (Phase II)
Goodness of Fit Computation

Step

0 Given the easiness estimates $\hat{\varepsilon}(i)$ of the items whose response fit is desired to be tested.

1 Specify the f(i,j) observed frequencies as defined by eqns (12) and (13) for those items as well as g(i,j) of eqn (14).

2 Compute the expected frequencies as given by (41) and (42).

3 Compute the χ^2 contribution as defined by (43) and (44).

4 Calculate the accumulated contributions of (45).

5 Select an α level for the significance test and find $\chi^2(1-\alpha, I-2)=\chi_o^2$.

6 Compute the test statistics, $\chi^2(k)$, according to (46).

7 If $\chi^2(k) \geq \chi_o^2$ for any k, reject H_o for those items and remove them from the analysis.

8 If any item is removed, re-estimate, with FRM (Phase I), the $\hat{\varepsilon}(i)$ of the remaining item and repeat steps 1-8 until no items are removed.

7. Estimation of Competency

We do not consider the estimation of the competency parameter of the BMM to be part of the FRM although it is a part of COMET. This arises because it is usually desired to estimate competency on examination responses resulting from administering tests composed of only a few, not all, of the available items. Since, at the time FRM is utilized these subsets have not been selected, the $\gamma(n)$ estimation process is delayed.

Following Wright and Panchapakesan (1969), Moonan and Covher (1973) have programmed a computer subroutine called "Maxco" to estimate $\gamma(n)$ from knowledge of total scores and the easiness of items on an examination. This process uses the Newton-Raphson iteration method to estimate $\gamma(n)$ corresponding to raw score ranging from 1 to k-1 in a test composed of k items. $\gamma(n)$ for raw scores 0 and K are non-estimable, the test being either too difficult or too easy for the examinee. We do know, however, that the competencies are either less than the competency associated with a raw score of 1 or greater than the competency associated with a raw score of k. The formula iterated in Maxco is

$$(47) \qquad f(\gamma(n)) = \sum_{k=1}^{k} \frac{\gamma(n) \cdot \hat{\varepsilon}(k)}{1 + \gamma(n) \cdot \varepsilon(k)} - r \qquad 0<r<k$$

where r is a raw score on a test composed of k items whose easiness are estimated by $\hat{\varepsilon}(k)$ and $\gamma(n)$ is the comptency of person n who obtained a raw score of r on the k items.

We have used Maxco on the I = K = 7 items of the FRM numerical example giving the results in Exhibit 5.

EXHIBIT 5

Competency Estimates for an Examination
of I = K = 7 ETST Items

Raw Score	Competency Estimate
0	non-estimable
1	.153
2	.382
3	.743
4	1.363
5	2.628
6	6.469
7	non-estimable

The standard errors of these estimates are not provided here.

8. The COMET System

The goal of Comet is to provide an integrated, systematic and objective procedure for the assessment of trainees in military programs which is based upon the Rasch Binary Measurement Model.

The entire system is composed of 12 interrelated sub-systems and the objectives of COMET are to identify, design and operationdize these sub-systems and integrate them into an effective systematic whole.

This paper's content is not intended to accomplish all of that but is intended as the first step, viz., the project definition and conceptual design phase. Actually it has done more than that since the development of the FRM, shown here, was a necessary sub-system without which the whole system would not be efficient or even feasible. We intend next to specify the essential sub-systems of COMET and to describe their functions. The following sub-systems have been identified

A Sub-System For	Brief Title
a. training needs assessment	Needs
b. transforming training needs into learning objectives	Objectives
c. training course design	Course
d. developing test items from learning objectives	Items
e. item response sampling	Sample
f. data collection	Data
g. data analysis	FRM
h. examination design	Test
i. test reproduction	Repro
j. competency estimation	Meas
k. evaluation	Eval
l. control	Cont

Description of Sub-System Functions

Needs

Planning the COMET system should start with the identification of training needs. A training need is defined as the discrepency between desired and required outcomes. The important notion is that to have a "need" we must identify and document that there is a gap between two outcomes, that which is currently resulting and that which should be resulting. A training specialist is probably required to design and operate this sub-system.

Objectives

It is convenient and perhaps necessary to transform identified needs into learning objective statements. There are several current means for doing this but none which the author cares to recommend. The training specialist should also be helpful in this sub-system.

Course

As was previously suggested, a training course design procedure needs to be followed to implement this sub-system. Probably no better source for this effort has been prepared other than that of Rundquist (1967).

Items

Having designed the training course and specified learning objectives we need to prepare items which, if answered correctly, will reflect trainee competency. This process may be significantly shortened by utilizing items made available from different services suggested by Wood (1968). Moonan (1972) provides a table of non-military behavioral objective and test item Pools ~~goals~~.

Sample

Having constructed the items for measuring competency we are required next to administer these items, perhaps in small batches, to individuals of varying training competency. As indicated earlier no formal sampling plan is required but responses may be acquired on a so-called "catch as catch can" basis.

Data

These response data may be recorded in various ways such as port-a-punch or mark sense cards or by an optical scanner form. An optical scanner system together with its mini-computer appears to be the optimal data collection and processing system for COMET. The cost effectiveness of this suggestion needs to be examined in any particular application.

FRM

Having collected the response data it needs to be organized in S(i,j) form and analyzed with the frequency ratio method, FRM in order to estimate the item easiness and to test the goodness of fit of the response data to the BMM. Items not fitting the model should be examined and purged from the item bank.

Test and Repro

A sub-system which can design automatically, prepare and duplicate competency examinations from the available item pool is required. This may be done via the computer or with the clerical system used by GLAKES. This process is sometimes referred to as CATC: Computer Assisted Test Construction.

Meas

Having specified the examination and the items therein we require the development of competency tables associated with possible raw scores on these tests. This is accomplished by using the Maxco subroutine developed by Moonan and Covher (1973).

Eval

Having made the measurements of training competency we can use this information in a variety of ways to make evaluations among which are the following:

a. Trainee competency

b. Change in trainee competency

c. Instructor capabilities

d. Training facility effectiveness

e. Training program effectiveness

f. Program objectives

A good discussion of these and other ways that the BMM can be used in evaluation is given in Hiscox (1974).

Cont

The COMET system needs to be controlled as a system by an effective manager who must constantly carry out the six functions of a manager: planning, directing, organizing, staffing, evaluating and controlling.

Any operational COMET system needs to be developed through systems analysis by an experienced or professional systems analyst.

9. Expected Implementation Problems

It is expected that the most difficult problem to be encountered in implementing COMET would be to find and utilize a systems' analyst to design the system for local conditions and constraints as well as to appoint a competent manager to monitor and manage the operational system. This person must possess many skills and have knowledge in the areas of measurement, computers, education, training, systems, etc. as well as the ability to work with specialists in those fields and with military personnel and systems. Such competent persons are difficult to locate.

Developing or finding items for the competency pool may be onerous but not too difficult. Interservice cooperation should be sought here and the MTA could play a significant and helpful role in these matters. A prototype COMET system might be designed and developed within one service and developmental costs shared. This system might then be used by other interested parties.

Standard errors of estimates of $\gamma(n)$ and $\varepsilon(i)$ are needed but not provided in this report. Their specification is under development but was not completed in time for inclusion in this paper. The science of examination design needs better formulation so that the test sub-system may make use of it. There appears to be only crude rules of thumb and theory available in psychometrics to assist us in this area..

All in all the practical problems do not appear too limiting, what is most needed is the appearance of a dynamic, resourceful, and persuasive leader to step forth from among our ranks and pursue the development and implementation of a COMET-like system within a training facility of his own service.

The major conceptual design work has been shown here, but there are many details yet to be worked out.

10. Summary

This report proposed, defined the scope and outlined the 12 subsystems of a general system called COMET whose general purpose is to objectively assess the competencies attained by trainees during the course of training in the military service. The measurement and evaluation features of COMET are based on a binary measurement model created by George Rasch. A new analysis procedure was developed to estimate a parameter of the model and to test the goodness of fit of responses of test items to the model. A numerical example of computations was provided and practical problems expected to be encountered during implementation are mentioned.

11. References

Blakely, J. A. (1969). Revision of ATAC System; Request for Proposal, Marine Corps Recruit Depot, San Diego.

Blommers, P. (1965). "A ratio Scale for Test Item Difficulty." Unpublished, University of Iowa, Department of Education.

Bramble, W. (1969). "A Least Square Method of Parameter Estimation for the Logistic Measurement Model." Unpublished, University of Kentucky.

Choppin, B. (1968). "Item Bank Using Sample-Free Calibration," *Nature*, August 24, p. 1-7.

Hatch, R. S. (1967). "Analysis of Automated Testing and Attrition Control Reports," Communication-Electronics Schools, MCRD, San Diego, CA.

Hiscox, S. B. (1974). "The Rasch Model in Evaluation," Paper presented at the Annual Meeting of NCME, Chicago, April, 1974.

Moonan, W. J. (1972). "Evaluation Trainee Test Performances with a G. Rasch Measurement Model," Proceedings of the 1972 MTA, San Antonio.

Moonan, W. J. (1972). "Handout for Presentation on Item Banking with the Rasch Model." Unpublished, NPTRL, San Diego, CA.

Moonan, W. J., and Covher, M. (1973). "MAXCO," Tools for Analysis and Decision, FACT DOCUMENT No. 24, NPTRL, San Diego, CA.

Rasch, G. (1960). Prohabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1969). "An Individual Centered Approach to Item Analysis with Two Categories of Answers." Unpublished, 1969 AERA Pre-Session on Rasch Models, Los Angeles, CA.

Rundquist, E. (1967). Course Design and Redesign Manual for Job Training Courses. San Diego, CA: NPTRL, SRR 66-17 (Revised).

Schmidt, W. (1969). "Rasch Model." Unpublished, 1969 AERA Pre-Session on Rasch Models, Los Angeles, CA.

Wheeler, W. H. (1966). "Learning Evaluation and Improvement Program." Unpublished, Electronics Technician Training School, NTC, Great Lakes, Illinois.

Wood, B. (1968). Educational Research, Vol. 10, No. 2.

Wright, B. (1968). Sample-Free Test Calibration and Person Measurement. *Proc. of the 1967 Invitational Conference on Testing Problems*. Educational Testing Service.

Wright, B. (1969). Lecture Notes. Unpublished, 1969 AERA Pre-Session on Rasch Models, Los Angeles, CA.

Wright, B., and Panchapakesan, N. (1969). "A Procedure for Sample-Free Item Analysis," *Educational and Psychological Measurement*, Vol. 29.

THE USE OF TACTICAL COMPUTERS TO PROVIDE WEAPONS AND TACTICS TRAINING TO COMBAT NCOs: RESULTS OF A FIELD TEST*

by

William G. Hoyt** and James D. Baker***

INTRODUCTION

Anticipating the smaller all volunteer force, the decentralization of training, and the added emphasis on individual and group instructional activities at the unit level, the Office of the Chief of Research and Development of the U.S. Army undertook an investigation in 1973 to determine whether the Army Tactical Data Systems (ARTADS) projected network of computerized systems could be utilized to support combat unit training activities. The study to be reported here was conducted in support of that effort. The objective of this research was to determine the extent to which Army tactical data systems could be used in a secondary, Automated Instruction (AI) role directed toward satisfying training requirements in a tactical unit environment. The test vehicle used for this assessment was the U.S. Army's Developmental Tactical Operations System (DEVTOS) which was located at Fort Hood, Texas. Detailed descriptions of the DEVTOS have been presented elsewhere(1,2,3)--suffice to say that it is a mobile automatic data processing (ADP) system which is intended to assist commanders and their staffs in the conduct of tactical operations by collecting, processing and summarizing information required for command decisions and staff action. Given the requirement and the system, what remained was to identify the subject matter to be taught; select the individuals who would take part in the study; develop the courseware necessary to teach it; develop a test plan for the conduct of the field test (which the Army designated as MASSTER Test FM 122, IBCS: Automated Instruction) and, finally, conduct the field test and analyze the results. The remainder of this paper will deal with the details of that undertaking.

---

* The opinions expressed are solely those of the authors and do not necessarily reflect official Department of the Army policy. Reproduction in whole or in part is permitted for any purpose of the United States Government.

** System Development Corporation, 2500 Colorado Avenue, Santa Monica, California 90406

*** U.S. Army Research Institute for the Behavioral and Social Sciences, 1300 Wilson Boulevard, Arlington, Virginia 22209, U.S.A.

APPROACH

Subject Matter and Courseware Considerations.

The maintenance of proficiency by MOS 11B40s (Light Weapons Infantry NCOs) was identified by the Combat Arms Training Board and the Continental Army Command (CONARC) as a significant unit training problem. Among the four subject matter areas required to be known by 11B40 personnel, Tactics and Crew Served Weapons were found to account for most of the proficiency test failures. For this reason, Crew Served Weapons and Tactics were the subject matter areas selected for AI courseware development. The other two (Individual Weapons and Field Activities) were not examined in the course of this study. Another area, General Educational Development (GED), was examined during the conduct of the study but the results of this aspect of the research will not be reported here. The training analysis (and subsequent recording of results) followed the principles and concepts contained in CONARC Regulation 350-100-1, Systems Engineering of Training. The data forms used to structure the AI courseware included Training Analysis Information Sheets, Criterion and Enabling Objectives, Test Items and Course Outlines. For crew served weapons, subject matter experts from the Weapons Committee of the Infantry School reviewed the training analysis results on a page by page basis, concurred (or non-concurred) in the topic selection, and then ranked each weapon as to its relative importance in the Crew Served Weapons Course. Three weapons ranked highest--LAW (light anti-tank weapon), 90 MM recoilless rifle and M60 machine gun. The resultant materials provided a basis for developing 12 hours of AI Crew Served Weapons courseware; although only four hours of the material was subsequently used in MASSTER Test 122. A similar procedure was used to provide 12 hours of AI courseware in the Tactics area.

AI Authoring Considerations.

The first task undertaken as part of this effort was to analyze the DEVTOS hardware and software to determine if it could support computer-aided instruction (CAI) and to survey and analyze existing computer-assisted instruction programs and procedures to determine their availability and feasibility for use within DEVTOS. The result of this phase of the effort is documented in detail elsewhere(4) but essentially it was found that the PLANIT(5,6,7,8) CAI system (PLANIT stands for Programming Language for Interactive Teaching) would best satisfy our requirements. Consequently, the courseware was written to run on the PLANIT/DEVTOS system.

In the development of course materials, every attempt was made to capitalize on the capabilities of PLANIT that are provided to assist with the author in preparing the instructional material and the student receive it. Among the presentation strategies used the most elementary was a straight instructional path. In addition, as the complexity of the topic warranted, an accelerated or a remedial instructional path was included. Courseware was written at a level designed to reduce reading difficulties. Military terms were included but where it appeared that they might introduce comprehension problems synonyms and/or examples were added. To the extent possible, on-line (on computer) representations of situations or examples were used. (Basically, if you can lay it out on a typewriter, you can put it on the scope). When this was neither possible nor feasible, pictures and diagrams were prepared as adjunct materials for use by the student.

Answer-matching for frames requiring constructed responses (PLANIT accepts certain misspelling of words) and the alternative selected for multiple choice questions constituted the student response data-base. Correct, as well as incorrect responses, provided a basis for selective remedial material and for making decisions for subsequent actions. Additionally, when an 11B40 answered incorrectly, he was not only told he was wrong, but why he was wrong. The student received feedback for each response entered. The feedback took a positive, negative or neutral form. Additional prompts and cues were also included. PLANIT has one other feature worthy of comment--it provides a course re-entry capability to restart the student at the point where he left off, in case he took a latrine or lunch break.

The PLANIT structured AI courseware content was reviewed by the Infantry School in July 1973. The subject matter experts comments ranged from "can go into practical application with little additional training (LAW)", to "this program looked good, the exercises on sighting and engaging targets were quite good and probably of the most value (90MM Rifle)". After a field tryout with ten 11 Bravo subjects course materials were finalized in August 1973.

Subject Selection.

The Test Plan (experimental design) called for 11B40 personnel who had scored relatively low on their MOS Proficiency Test; ergo, they needed the training. To get this data a run was made on the personnel tapes at Ft. Hood and the 11B40 files were extracted from it. Card decks containing names and Social Security Numbers of 11B40 personnel were sent to the Enlisted Evaluation Center, Fort Benjamin Harrison, Indiana and the 1972 MOS Proficiency Test scores were obtained. A month before MASSTER Test 122, the card decks were again run against the personnel tapes to determine whether the 11B40s were still at Fort Hood. On the basis of this information, listings of the subject pool of 11B40s were prepared and delivered to Headquarters MASSTER. For each 11B40 on the list, we had considerable background data including his GT (general technical) score, MOS proficiency test scores, education, age and pay-grade on a student record form.

Conduct of the Field Test.

MASSTER Test FM 122 commenced on 4 September 1973. An outline of the test design in terms of the treatment of participants is shown in Table 1. Briefly, the test utilized 120 subjects; 60 in the AI Group, 30 in the Study Group and 30 in the Control Group. One-half of the Ss in each of the AI and Study Groups participated in one of the two subject matter areas, Crew Served Weapons or Tactics.

The procedure was for the 11B40 subjects (maximum of 12 each day) to fill out an introductory form on arrival and be briefed on the purpose of MASSTER Test FM 122. While the briefing was conducted, student records were pulled and assignments made to either Crew Served Weapons or Tactics based upon MOS proficiency test score. For a given S his past scores were compared and he was assigned to that treatment group which corresponded to his lowest, previous proficiency test score. Ss then took their Pre-test (half on Form A and half on Form B) and were assigned at random to the AI, Study or Control Groups. The AI Group took their course (e.g., Tactics) on the computer; the Study Group studied the same material from Army publications, and the Control Group had an unrelated assignment, learning the ALPHA-DOT System of communication.(9,10) Each group averaged approximately four hours on their activity. The Post-test (the alternate form of the pre-test) followed, and then the AI Group was interviewed.

RESULTS

In discussing results, there is always the question as to whether one or more groups were favored (biased) in regard to background variables or pre-test scores and pre-test time. Shown in Table 2 are the means and standard deviations for the AI, Study and Control Groups for the LAW (light anti-tank weapon). The data show that the three groups are about the same which indicates the effects of these variables were virtually cancelled out by the random assignment of test subjects to the AI, Study and Control Groups. The one variable that has the greatest difference is GT score where the Control Group has a mean of 108, while the AI and Study Groups have means near 100. This difference is not regarded as having a serious effect since the correlation of GT score and pre-test score was low, .26, and the difference was in favor of the Control Group. Table 2 also shows that we have a fairly wide range of 11B40 personnel represented. In the AI Group, for example, ages were from 21 to 45; education from 8th grade to 13th; GT scores from 88 to 119; and Paygrade from 5 to 7.

Although the mean pre-test times (variable 21, Table 2) for the three groups were within a 1/2 minute of each other, the mean post test of the topic warranted, an accelerated or a remedial instructional path was included. Courseware was written at a level designed to reduce reading difficulties. Military terms were included but where it appeared that they might introduce comprehension problems synonyms and/or examples were added. To the extent possible, on-line (on computer) representations of situations or examples were used. (Basically, if you can lay it out on a typewriter, you can put it on the scope). When this was neither possible nor feasible, pictures and diagrams were prepared as adjunct materials for use by the student.

Answer-matching for frames requiring constructed responses (PLANIT accepts certain misspelling of words) and the alternative selected for multiple choice questions constituted the student response data-base. Correct, as well as incorrect responses, provided a basis for selective remedial material and for making decisions for subsequent actions. Additionally, when an 11B40 answered incorrectly, he was not only told he was wrong, but why he was wrong. The student received feedback for each response entered. The feedback took a positive, negative or neutral form. Additional prompts and cues were also included. PLANIT has one other feature worthy of comment--it provides a course re-entry capability to restart the student at the point where he left off, in case he took a latrine or lunch break.

The PLANIT structured AI courseware content was reviewed by the Infantry School in July 1973. The subject matter experts comments ranged from "can go into practical application with little additional training (LAW)", to "this program looked good, the exercises on sighting and engaging targets were quite good and probably of the most value (90MM Rifle)". After a field tryout with ten 11 Bravo subjects course materials were finalized in August 1973.

Subject Selection.

The Test Plan (experimental design) called for 11B40 personnel who had scored relatively low on their MOS Proficiency Test; ergo, they needed the training. To get this data a run was made on the personnel tapes at Ft. Hood and the 11B40 files were extracted from it. Card decks containing names and Social Security Numbers of 11B40 personnel were sent to the Enlisted Evaluation Center, Fort Benjamin Harrison, Indiana and the 1972 MOS Proficiency Test scores were obtained. A month before MASSTER Test 122, the card decks were again run against the personnel tapes to determine whether the 11B40s were still at Fort Hood. On the basis of this information, listings of the subject pool of 11B40s were prepared and delivered to Headquarters MASSTER. For each 11B40 on the list, we had considerable background data including his GT (general technical) score, MOS proficiency test scores, education, age and pay-grade on a student record form.

Conduct of the Field Test.

MASSTER Test FM 122 commenced on 4 September 1973. An outline of the test design in terms of the treatment of participants is shown in Table 1. Briefly, the test utilized 120 subjects; 60 in the AI Group, 30 in the Study Group and 30 in the Control Group. One-half of the Ss in each of the AI and Study Groups participated in one of the two subject matter areas, Crew Served Weapons or Tactics.

The procedure was for the 11B40 subjects (maximum of 12 each day) to fill out an introductory form on arrival and be briefed on the purpose of MASSTER Test FM 122. While the briefing was conducted, student records were pulled and assignments made to either Crew Served Weapons or Tactics based upon MOS proficiency test score. For a given S his past scores were compared and he was assigned to that treatment group which corresponded to his lowest, previous proficiency test score. Ss then took their Pre-test (half on Form A and half on Form B) and were assigned at random to the AI, Study or Control Groups. The AI Group took their course (e.g., Tactics) on the computer; the Study Group studied the same material from Army publications, and the Control Group had an unrelated assignment, learning the ALPHA-DOT System of communication.(9,10) Each group averaged approximately four hours on their activity. The Post-test (the alternate form of the pre-test) followed, and then the AI Group was interviewed.

RESULTS

In discussing results, there is always the question as to whether one or more groups were favored (biased) in regard to background variables or pre-test scores and pre-test time. Shown in Table 2 are the means and standard deviations for the AI, Study and Control Groups for the LAW (light anti-tank weapon). The data show that the three groups are about the same which indicates the effects of these variables were virtually cancelled out by the random assignment of test subjects to the AI, Study and Control Groups. The one variable that has the greatest difference is GT score where the Control Group has a mean of 108, while the AI and Study Groups have means near 100. This difference is not regarded as having a serious effect since the correlation of GT score and pre-test score was low, .26, and the difference was in favor of the Control Group. Table 2 also shows that we have a fairly wide range of 11B40 personnel represented. In the AI Group, for example, ages were from 21 to 45; education from 8th grade to 13th; GT scores from 88 to 119; and Paygrade from 5 to 7.

Although the mean pre-test times (variable 21, Table 2) for the three groups were within a 1/2 minute of each other, the mean post test

TABLE 1

TEST DESIGN--TREATMENT OF PARTICIPANTS

AI GROUP (N=60)	STUDY GROUP (N=30)	CONTROL GROUP (N=30)
PRE-TEST	PRE-TEST	PRE-TEST
AUTOMATED INSTRUCTION	SELF STUDY	ACTIVITY UNRELATED TO SUBJECT MATTER
POST-TEST	POST-TEST	POST-TEST
STRUCTURED INTERVIEW		

CREW SERVED WEAPONS: GROUP MEANS AND STANDARD DEVIATIONS

VARIABLE NAME (NO.)	AI GROUP (n=33)		STUDY GROUP (n=13)		CONTROL GROUP (n=13)	
	M	SD	M	SD	M	SD
GT SCORE (26)	100.2	8.9	100.7	11.0	108.3	12.8
EDUCATION (4)	12.3	1.6	12.1	1.8	12.1	2.1
AGE (5)	28.4	5.6	30.9	6.0	29.6	5.3
PAYGRADE (2)	5.7	0.6	5.9	0.5	5.8	0.6
MOS TEST 2 (10)	13.5	2.8	14.2	2.6	12.7	2.9
MOS TEST TOTAL (13)	65.9	8.6	66.1	8.4	64.2	10.5
PRETEST SCORE (6)	10.3	2.7	10.1	2.5	10.4	3.1
PRETEST TIME (21)	25.6	6.3	26.2	4.9	25.6	7.6
POSTTEST SCORE (7)	18.7	3.5	15.6	3.1	12.6	3.1
POSTTEST TIMES (23)	15.0	5.2	22.9	4.9	20.5	8.3
GAIN SCORE (25)	8.4	3.1	5.5	3.8	2.2	3.0

TABLE 2

time (variable 23) differed considerably. The Study Group took 8 minutes or 53% longer to complete the post test than the AI Group. Based upon feedback obtained during the interview, this indicates that the AI Group was more confident of their knowledge and skills than the Study Group and were able to answer the test questions more quickly and as we shall see, more accurately.

Between Groups Comparison.

The two critical comparisons for the study are between: (1) The AI and Control Groups and (2) The AI and Study Groups. Statistical summaries for the LAW findings are shown in Table 3. The AI Group had a mean gain score of 8.394, an 82% increase over their pre-test score. The Control Group had a mean gain score of 2.154, a 21% increase. The difference between the two mean gain scores is 6.240 which, using the T-test of statistical significance, produces a T ratio of 6.23 which is highly significant at the .01 level (.01 = 2.69). The T ratio shows that the possibility of a mean difference 6.240 occurring by chance is remote. Consequently, this difference can be attributed to the automated instruction given the AI Group. THE SIGNIFICANT T RATIO AND THE 82% INCREASE IN PROFICIENCY ARE POSITIVE STATISTICAL AND PRACTICAL EVIDENCE THAT LEARNING TAKES PLACE WHEN TACTICAL DATA SYSTEMS ARE EMPLOYED IN A SECONDARY (AUTOMATED INSTRUCTION) ROLE.

The same comparisons were made between the AI and Study Groups. The difference was statistically significant at the .05 level, producing a T ratio of 2.66 in favor of the AI Group. THE SIGNIFICANT T RATIO AND A 52% INCREASE IN PROFICIENCY OVER THE STUDY GROUP ($\frac{82\% - 54\%}{54\%} = 52\%$) ARE POSITIVE STATISTICAL AND PRACTICAL EVIDENCE THAT--WITHIN THE LIMITS OF THIS STUDY--LEARNING BY MEANS OF AUTOMATED INSTRUCTION IS MORE EFFECTIVE THAN STUDY GROUP METHODS OF TRAINING. For purposes of completeness it should be noted that the differences between the Study Group and Control Group produced a T ratio of 2.54 which is significant at the .05 level. Thus the Study Group also had a significant gain in learning when compared to the Control Group, although not as great as that of the AI Group. The relationship of these between group findings are graphically portrayed in Figure 1

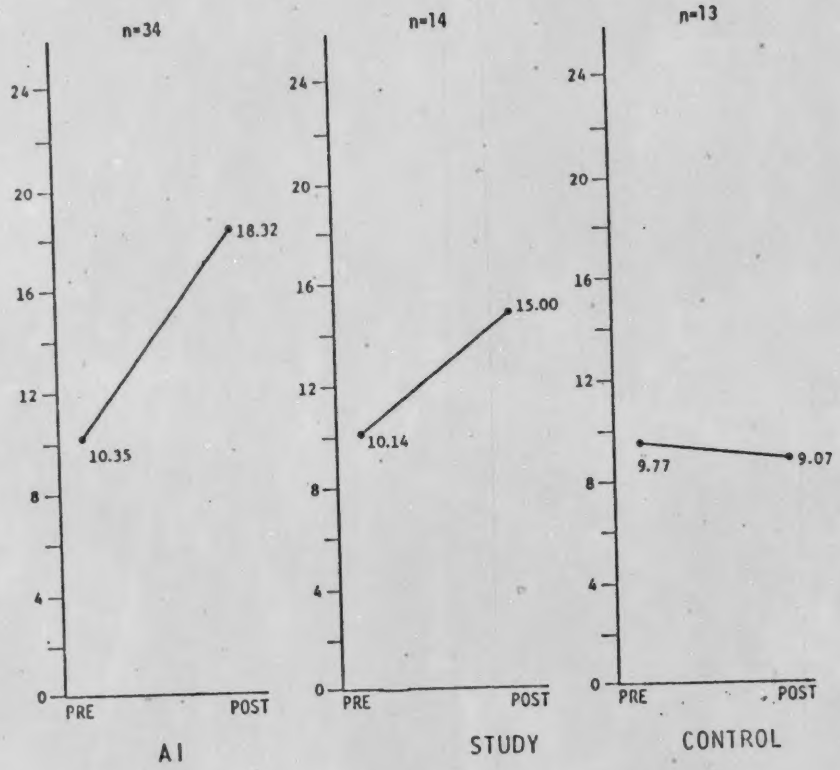
The Study Group situation was better in this experiment than the 11B40 normally encounters. He was given a pre-test and while he did not know his pre-test score, he had some idea of how well he did on the test. He had the field manuals to work with plus an instruction sheet telling him what topics to cover and in what order, and on what pages and paragraphs they were located. Finally he had additional, supplemental material available in this structured-study situation.

The results for the LAW were quite similar to those for Tactics, as evidenced by the data presented in Figure 2. There is some difference in the slope of the control group curve from that obtained for the Tatics

RESULTS OF THE CREW SERVED WEAPONS STUDY

TREATMENT GROUP	n	MEAN PRETEST SCORE (VAR. 6)	MEAN POSTTEST SCORE (VAR. 7)	PERCENT INCREASE PRETEST TO POSTTEST	MEAN GAIN SCORE (VAR. 25)	GAIN SCORE STANDARD DEVIATION
AI	33	10.273	18.667	82%	8.394	3.082
S	13	10.077	15.615	54%	5.538	3.755
C	13	10.384	12.538	21%	2.154	2.996

TABLE 3



TACTICS

Figure 1

group. But this pre- and post-test difference shown for the LAW control group was not statistically significant and appears to reflect the variability between the alternative forms of the test. It also merits comment that the "idealized" equal-cell entries sought in our original Test Plan design (Table 1) were not fully obtained in reality. These differences in "idealized" and "actual" may be obtained by comparing the Ns of Figures 1 and 2 to those given in Table 1. Such deviations from the planned (or how the field test differs from the laboratory) are to be expected in this type of setting and situations such as this has been the topic of several recent articles(11,12).

Relationship of Data to GT (General Technical) Score.

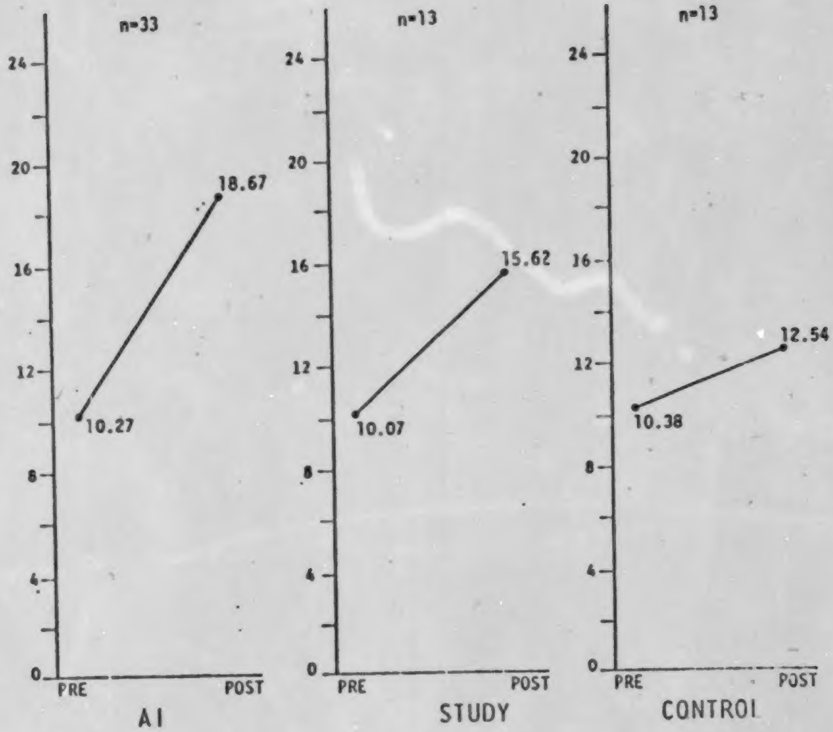
A total of 77 variables were analyzed for the 33 subjects in the Crew Served Weapons AI Group and a similar number for the Tactics AI Group. The statistics included the frequency distributions, means, standard deviations, and range of scores for each variable; and the intercorrelation matrix for all 77. The relationship of four specific variables to GT score is shown in Figure 3. Pre-test scores are shown by the Xs; Post test by the Os and the gain score by the line between. GT score is shown along the abscissa.

Of particular interest to the training community is GT score which is considered a measure of general aptitude or ability to learn. Combat arms personnel, many of whom are in the lower ranges of GT score, are considered to present special problems in training for the military services. As shown here, GT scores for the 33 11B40s in the Crew Served Weapons AI Group ranged from 88 to 119 with 55% of them below 100. THESE RESULTS SHOW THAT THE AUTOMATED INSTRUCTION METHOD OF TRAINING APPLIES EQUALLY WELL TO 11B40 PERSONNEL WITH HIGH OR LOW GT SCORES. The 10 AI subjects with the lowest GT scores had an average post test score of 18.7 and the 10 highest, an average of 19.5, less than a point difference. Similar results were obtained for the Tactics AI Group.

These RESULTS INDICATE THAT AUTOMATED INSTRUCTION IS AN EFFECTIVE METHOD OF PROVIDING WEAPONS AND TACTICS TRAINING ACROSS THE RANGE OF 11B40 GT SCORES. Automated instruction apparently has the effect of reducing or overcoming the verbal handicaps usually associated with lower GT scores.

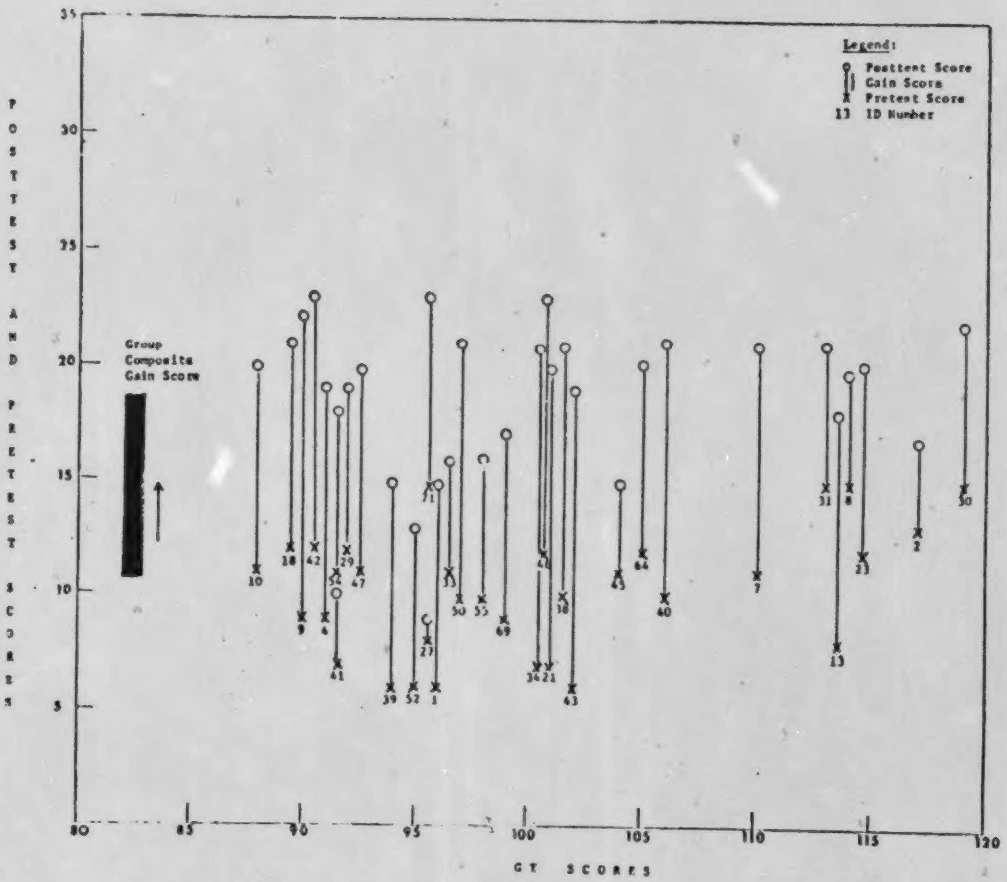
Other Variables Examined.

The three other variables--Paygrade, Education and Age--were examined and showed a low negative correlation with gain score. Thus AUTOMATED INSTRUCTION APPEARS TO BE EFFECTIVE ACROSS AGE GROUPS, ACROSS EDUCATION LEVEL, AND ACROSS THE PAYGRADE OF THE 11B40 POPULATION.


n=33
24
20
16
12
8
4
0
18.67
10.27
PRE
POST
AI
n=13
24
20
16
12
8
4
0
15.62
10.07
PRE
POST
STUDY
n=13
24
20
16
12
8
4
0
12.54
10.38
PRE
POST
CONTROL

LAW

Figure 2



Relationship of Pretest, Posttest and Gain Scores (Var. 25) to GT Scores (Var. 26) for CSW AI Subjects (n = 33)(Correlation between Gain and GT is -.14)

Figure 3

The total course time for the 33 subjects in the LAW course averaged 224 minutes and ranged from 158 minutes to 300 minutes. There were 4 lessons in the LAW course. Twenty-four of the 33 subjects completed the entire LAW course; one was in lesson 4 at the time the test period ended and 8 had completed or were in lesson 3.

A ratio score was obtained to determine whether slow learners used a different strategy in taking the AI course than did the faster learners. The total entries each slow learner made was divided by the total FL Frames (the minimum path to the point he reached in the course). This ratio is the number of entries made for each FL frame. THE RESULTS SHOWED THAT THE SLOWEST 16 MADE ONLY ONE ADDITIONAL ENTRY MORE THAN THE FASTEST 16 PER EVERY NINE FL FRAMES REACHED. THIS RELATIVELY SMALL DIFFERENCE WOULD INDICATE THAT THE SLOWER LEARNERS WENT THROUGH THE AI COURSE IN THE SAME WAY AS THE FAST LEARNERS AND SIMPLY REQUIRED MORE TIME TO READ AND COMPREHEND THE MATERIAL.

Qualitative Results.

After the post test, subjects in the AI Group were interviewed in depth with regard to their experience with Automated Instruction. In the conduct of the interview the interviewer filled out a form as the subject responded. Some questions were open ended, while others required a specific answer. The interviews were also taped--with the S's permission. The response to and acceptance of automated instruction by 11B40s was striking. The idea that combat infantry NCOs, who start fidgeting after an hour in the classroom, would sit down at a scope for four or five hours and become completely absorbed in the learning process is indicative of the power of this media. As one qualified NCO drill instructor aptly expressed it: "It covered the whole weapon. When you come out of there, you knew a hell of a lot more than you did when you went in." Another NCO expressed it this way: "Makes you confident because when you walk out, you know the subject."

The responses to the interview questions were virtually unanimous. In terms of the expected, and widely touted, positive features of CAI, these Ss too remarked on the quiet atmosphere, self-pacing features, immediate feedback, individual attention and non-threatening interactions between the student and the computer. The latter point was a key feature since these NCOs repeatedly remarked that they were loathe to ask questions or make mistakes in a classroom for fear of "looking stupid" in the eyes of those they command. Regarding the course materials and technique of presentation, the Ss commented that the items were easy to understand; obviously built on each preceeding item; were accurate, and "got at the facts without the B.S." From the student's standpoint he found the situation to be challenging (but non-threatening), rewarding because you always know immediately where you stand and he believed that new methods of training such as AI would make Army instruction better and more interesting.

SUMMARY AND CONCLUSIONS

With the present efforts underway in the U.S. Army to develop and field tactical data systems, it is likely that there will be a considerable data processing capability at the tactical unit level in the near future. A potential secondary role for these systems, when they are not required for tactical operations, is that of supporting unit and individual training. The results from the U.S. Army's MASSTER Test FM 122, "IBCS: Automated Instruction" has demonstrated the feasibility of using these systems in a stand alone mode in support of softskills unit training requirements--in this case MOS and GED training of infantry personnel.

Based upon our experience with this study we may conclude that one objective of the National Science Foundation (NSF) effort in the development of PLANIT--maximum portability--has been validated, We installed it on a DEVTOS tactical data system and it is presently in the process of being installed on a next generation, militarized and ruggedized tactical data system. Additionally, a fair number of diverse courseware packages were developed and successfully executed on a tactical computer using this CAI language.

We may further conclude that automated instruction in a field setting is effective, is enthusiastically accepted by NCOs, and is more efficient than the traditional study-method of training. AI training has the advantage of reducing or overcoming the verbal handicaps usually associated with lower GT scores. Even for these "atypical Ss" (the wags would have us believe that psychology is the study of college sophomores, while CAI is the study of Palo Alto school children) automated instruction was found to hold their attention, required them to think about what they were doing, and patiently provided them the time in which to ponder and learn. Further, it provided a positive learning experience in a non-threatening environment.

As all behavioral science studies must (or seem to), this study too ends on the note that based upon our findings further research seems warranted. A number of potential studies come to mind, but one in particular stands out. Over and over again in our debriefing of the LAW students the point was made that, based upon their AI experience, they felt they could have left the console and gone out and actually fired the weapon with real accuracy. In this regard, there has been of late considerable interest in restructuring the criteria against which tests are designed and constructed, viz criterion-referenced testing.(13)

Criterion-referenced tests are said to be designed "to yield measurements that are directly interpretable in terms of specified performance standards."(14) Traditionally developed tests (such as achievement scored, norm-referenced proficiency tests) are usually the means used for assessing such things as the efficacy of using AI as the instructional media. Such was the case in the present study. But the real experimental question remains: Does CAI teach effectively enough in the abstract to impact on performance which is being measured in the concrete? Current plans call for testing this notion at the U.S. Army Infantry School, in Fort Benning, Georgia.

In closing, heraldry has always been an integral part of the military milieu. In the case of this field test, we had a wall poster which we used in conjunction with our experiment as an unofficial coat of arms. Inscribed on it was a motto, whose source we do not know, but which stated succinctly our feelings throughout: "Give me a fish and I eat for a day--teach me to fish and I eat for a lifetime."

REFERENCES

1. Baker, J.D. Human factors experimentation within a tactical operations system (TOS) environment, BESRL Research Study 68-4, October 1968.

2. Baker, J.D. Acorns in flower pots/psychologists in the field. JSAS Catalog of Selected Documents in Psychology, 1972, 2; p 88.

3. Baker, J.D. Human performance research in miliary information systems: Quo fuimus et quo vadimus? Proceedings: NATO Defense Research Group Seminar on Man-Machine Relations, May 1973.

4. Hoyt, W.G., Butler, A.K. and Bennik, F.D. Application of tactical data systems for training: Volume II - AI/DEVTOS automation studies. ARI Research Memorandum 74-8, July 1974.

5. Frye, C.H. CAI languages: Their capabilities and applications. Datamation, September 1968.

6. Feingold, S.L. PLANIT - a language for CAI, Datamation, September 1968.

7. Frederick, F.J. PLANIT: A "machine-independent" CAI language and system. Viewpoints (Indiana University), Vol. 50, No.4, July 1974.

8. Atkinson, R. and Wilson, H.A. (Eds.) Computer-Assisted Instruction: A book of readings, Academic Press: New York, 1969.

9. Sidorsky, R.C. Alpha-Dot: A new approach to direct computer entry of battlefield data. ARI Technical Paper 249, January 1974.

10. Sidorsky, R.C. Source data automation via the alpha-dot tablet: A feasibility study. ARI Technical Paper, in press. (Abbreviated version presented at 82nd Annual Convention, American Psychological Association, September 1974).

11. Johnson, E.M. and Baker, J.D. Field testing: The delicate Compromise. Human Factors, 3, June 1974.

12. Special Issue: Field testing: Human Factors, 3, June 1974.

13. Carver, R.P. Two dimensions of tests: Psychometric and edumetric. American Psychologist, 7, July 1974.

14. Glaser, R. and Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.) Educational measurement. Washington, D.C.: American Council on Education, 1971.

SECTION XII

PERSONNEL ASSESSMENT

ASSESSMENT AND COUNSELING FOR LEADERSHIP DEVELOPMENT

J. A. Caviness, Ph.D.
Principal Scientist, Fort Benning R&D Field Unit
US Army Research Institute for the Behavioral and Social Sciences

INTRODUCTION

In February 1973, the Army Chief of Staff directed that an Assessment Center be established at Fort Benning. The Center is a pilot project to test the applicability of personnel selection and leadership development techniques.

Assessment center techniques have been used successfully as both a selective device and in professional self-development by other armed forces, including the British, Australian, and Israeli armies, by governmental agencies, and by civilian industries.

The Assessment Center evaluates leadership behavior using situational exercises which simulate "on-the-job" conditions. An individual is confronted with situations that he could face if he were promoted or selected for commissioning. Trained assessors develop a profile of the individual's effectiveness in his current grade and his potential effectiveness at a higher level. This profile, plus video recordings of the man's performance, can be used in a counseling session. This counseling is aimed at professional self-development and is focused on the relative strengths and weaknesses of each individual.

The mission of the United States Army Assessment Center (ACTR) is to:

a. Test the assessment concept as a precommissioning screening device.

b. Test the assessment concept as a leadership development technique.

c. Determine whether the concepts and techniques developed are feasible and applicable for Army-wide use.

To accomplish this mission, ACTR was organized at the directorate level within the Infantry School. Sponsorship is provided by DCSPER. DCSPER provides guidance and is the Department of the Army point of contact. Guidance is also provided by TRADOC. The Army Research Institute was instrumental in conceptualizing and implementing the assessment program, and is conducting research that will provide data for the evaluation of the project.

DESCRIPTION

ACTR focuses on job functional leadership behaviors, those that mark superior leading. (Negative behaviors that mark failure or inferiority are not stressed.) The twelve dimensions of leader behavior are:

adaptability
administrative skills
communication skills
decision making
forcefulness
mental ability
motivation
organizational leadership
physical fitness
social skills
supervisory skills
technical & tactical competence

ACTR assesses these dimensions of leader behavior using nine individual and group exercises.

Group Exercises:

Leaderless Group Discussion
Competitive Stock Trading Exercise
Rotating Assigned Leadership Exercise

Individual Exercises:

Entry Interview by Assessor
Interview of Assessee by Assessee
In-Basket Exercise
Writing Exercise
Organizational Simulate
Physical Training Exercise

In addition to the individual and group exercises, the following psychometric tests are used as an integral part of assessment:

Leadership Opinion Questionnaire
The Leadership Q-Sort Test
Watson-Glaser Critical Thinking Appraisal
The Henmon-Nelson Test of Mental Ability
The Nelson-Denny Reading Test
The Social Insight Test (Chapin)
Work Environment Preference Schedule (Gordon)
Strong Vocational Interest Blank for Men
Edwards Personal Preference Schedule

OBJECTIVES

The objectives of assessment were two: first, to use the assessment process as a source of information that would be used as a basis for counseling aimed at leadership development; second, to use the assessment process as a source of information that would be used as a basis for screening candidates for commissioning. In short, ACTR was in fact two centers, one focusing on professional development, the other focusing on selection (or screening) for commissioning.

The use of the assessment processs for professional development was tested on two populations: students entering the Infantry Officer's Advanced Course (IOAC) and the students entering the Advanced Non-Commissioned

Officer's Educational System (ANCOES). Because these two populations have different professional experiences and organizational expectations, the assessment exercises developed were different (although parallel).

The use of the assessment process for screening candidates for commissioning was tested on the students entering the Branch Immaterial Officer Candidate Course (BIOCC), and was expanded to include students entering the Infantry Officer Basic Course (IOBC).

The objectives, screening and development, were two; the populations sampled were four:

Selection:

Branch Immaterial Officer Candidate Course (8 groups)
Infantry Officer's Basic Course (6 groups)

Development:

Advanced Non-Commissioned Officer's Educational System (5 groups)
Infantry Officer's Advanced Course (5 groups)

PROCEDURE

DEVELOPMENT

The developmental activities, conducted during the second half of FY 73, consisted of 1) selection of behavioral dimensions to be assessed, 2) development of assessment exercises and questionnaires, 3) selection of psychometric tests, 4) training of staff in assessing and counseling, 5) development of data processing techniques and 6) pilot testing the assessment process.

OPERATION

The second phase of the project involved assessing 24 groups of assessees. Each group comprised (usually) eighteen men. Ten of the groups were counseled. This phase occupied FY 74 and the first half of FY 75.

FOLLOW-UP

The students from BIOCC, IOBC, IOAC, and ANCOES who attended ACTR for evaluation and/or counseling and their classmates who were designated as control subjects are being studied in a follow-up research program for two years after assessment.

The data collection in the follow-up research program includes:

1) asking the assessees how they profitted from attendance at ACTR, and

2) collecting comparative data on assessees and control classmates.

The schematic timetable for the follow-up research program is:

1) Upon graduation from the relevant course (BIOCC, IOBC, IOAC, or ANCOES) data is collected on class standing, and where available, leadership.

2) Six months after graduation from the relevant course (IOBC, IOAC, BIOCC, or ANCOES) and entry into an assignment, a set of questionnaires is used to collect data on the effects of assessment and counseling, when given, on performance on the job.

3) Eighteen months after graduation the above (2) data collection will be repeated. In addition, OERs and EERs will be examined for data pertaining to on-the-job proficiency and promotion records will be examined for data pertaining to career advancement.

PRESENT PAPER

The present paper addresses the question: Does the assessment/counseling program (for ANCOES and IOAC) result in a change in self-opinion? Does he learn something about himself?

The assessment exercises, implicitly and by induction, illuminate objectives or standards of performance. The assessees are exposed to and become aware of what is expected of them. Peer performance clearly demonstrates the inherent value of striving to meet the requisite standards of behavior. The test of the effectiveness of entering into the spirit of doing one's best is empirical. Verification of these perceptions of objectives and performance is given in the counseling sessions, where the main goals are making the implicit explicit and in showing that the specific lessons learned (at ACTR) have generalizibility to the real world.

It will be the goal of the follow-up study to show how much is learned in the long run. That there is a flash of insight is clear. Whether or not the flash stimulates an increase in self-development or self-improvement, and whether there is continued improvement over time, time and data will tell.

The first step in the data analysis can offer a tentative answer to the question: Does the ACTR program generate a change in self-concept? Both before entry into ACTR and after counseling at ACTR the following self-rating scales are administered:

1) Skills in getting along with co-workers (Social)

2) Skills in getting ideas across face-to-face (Oral)

3) Skills in getting ideas across in writing (Writing)

4) Skills in leading and motivating others (Leadership)

5) Skills in dealing with the administrative procedures (Administrative)

6) Skills in understanding problems and figuring out answers (Problem Solving)

Simply put, the question addressed in this preliminary analysis is, does the assessment/counseling process result in a shift on the six scales?

Since information relating to these scales is available from the assessor ratings, it is also possible to ask whether self-ratings at ACTR are significantly related to assessor ratings at ACTR.

RESULTS

SELF-RATINGS (t for paired scores)

The data are strikingly clear. In both ANCOES and IOAC groups, the shift on each of the six scales is significant and downward (see Table 1).

RELATIONSHIPS BETWEEN SELF-RATINGS AND ASSESSOR RATINGS (Pearson Product Moment)

The relationships between self-rating and assessor ratings are shown in Table 2. At entry, the data indicate significant relationships in one case for IOAC and in no cases for ANCOES. After counseling, the data indicate significant relationships in three cases for IOAC assessees and four cases for ANCOES.

DISCUSSION

The shifts in self-ratings are offered as evidence of the effect that the assessment/counseling experience has on self-opinion. When the assessee enter ACTR, his self-rating scores, on the average, are higher than when he leaves ACTR. He has changed his opinion of himself. He has learned about himself. He has been assessed, enlightened, educated.

It is probable that the downward shift in self-ratings results from the perception that the assessee has of himself as he performs in the assessment exercise. It is possible that being counseled also contributes to the change in self-perception. If an assessee is scored low by the assessors and in accordingly counseled, his shift downward might be expected to be greater than that of as assessee who is scored high by the assessors and accordingly counseled. (The latter assessee might shift his self-rating upward.)

Assessor ratings and self-ratings at ACTR are not perfectly related. Each rating has its own value in that it reflects a view, but the relationships between the ratings have a special value. The relationships between assessee self-ratings at entry (before the assessment program begins) and assessor ratings (composite scores representing the cumulative observations of the several assessors over the several exercises) indicate that only the ratings by and of the IOAC assessees on Problem Solving are in agreement (in terms of systematically varying directly proportionately with each other). After counseling, the agreement is better. A total of seven significant relationships is found. (Oddly enough, the earlier relationship between ratings on Problem Solving for IOAC assessees suffers a diminution.) It is likely that the ACTR assessment/counseling program plays a role in effecting the change by the assessees towards better agreement with the assessors, especially in those skill areas most sharply tested in the highly verbal and interpersonal assessment context.

The assessment/counseling program does result in a change in self-concept. The assessee learns about himself by experiencing his performance in the exercises. In addition, the assessee is taught about himself by a counselor who has the results of the several assessor reports. The lessons come quickly, in a briskly paced three-day schedule, at first implicitly and by induction, but finally explicitly and by example in a three-hour counseling session. In the counseling session, the assessee sees and hears himself (via video tape) in excerpts from the exercises; the clarity with which the TV brings home the message is startling.

TABLE 1

Comparison of Self Ratings Before and After Assessment and Counseling

	Self-Rating At Entry		Self-Rating After Counseling		
	$\bar{X}$	SD	$\bar{X}$	SD	t
ANCOES (N=87)					
Social	3.71	.50	3.08	.97	6.18**
Oral	3.36	.68	2.84	.90	4.89**
Writing	2.59	.86	2.07	.97	4.48**
Leadership	3.38	.67	2.97	.88	3.97**
Administrative	2.85	.81	2.53	.97	2.80**
Problem Solving	3.32	.62	3.01	.88	2.88**
IOAC (N=88)					
Social	3.49	.69	3.01	.86	4.92**
Oral	3.15	.65	2.92	.86	2.58*
Writing	2.53	.92	2.17	1.00	3.69**
Leadership	3.25	.67	3.01	.72	3.48**
Administrative	2.99	.77	2.57	.80	4.24**
Problem Solving	3.11	.67	2.82	.74	3.57**

* Significant at .05 level
** Significant at .01 level

TABLE 2

Comparison of Assessor Ratings with Self Ratings Before Assessment and Counseling and with Self Ratings after Assessment and Counseling

	Assessor Ratings at ACTR		Self-Rating at Entry			Self-Rating After Counseling		
	X	SD	$\overline{X}$	SD	r	$\overline{X}$	SD	r
ANCOES (N=87)								
Social	3.25	.47	3.71	.50	+.08	3.08	.97	+.15
Oral	3.20	.55	3.36	.68	+.05	2.84	.90	+.30**
Writing	2.84	.77	2.59	.86	+.18	2.07	.97	+.45**
Leadership	1.89	.62	3.38	.67	-.11	2.97	.88	+.13
Administrative	2.67	.61	2.85	.81	+.14	2.53	.97	+.27*
Problem Solving	3.11	.57	3.32	.62	+.06	3.01	.88	+.23*
IOAC (N=88)								
Social	3.14	.53	3.49	.69	+.16	3.01	.86	+.23*
Oral	3.47	.49	3.15	.65	+.16	2.92	.86	+.42**
Writing	3.25	.69	2.53	.92	+.15	2.17	.99	+.52**
Leadership	2.33	.37	3.25	.67	-.04	3.01	.72	+.03
Administrative	2.60	.52	2.99	.77	-.05	2.56	.80	-.06
Problem Solving	3.28	.64	3.11	.67	+.26*	2.81	.74	+.16

* Significant at .05 level
** Significant at .01 level

ASSESSMENT AND COUNSELING FOR LEADERSHIP DEVELOPMENT

J. A. Caviness, Ph.D.
Principal Scientist, Fort Benning R&D Field Activity
US Army Research Institute for the Behavioral and Social Sciences

-ABSTRACT-

In February 1973, the Army Chief of Staff directed that an Assessment Center (ACTR) be established at Fort Benning to test the applicability of personnel selection and leadership development techniques.

Description: ACTR focuses on job functional leadership behaviors, those that mark superior leading. (Negative behaviors that mark failure or inferiority are not stressed.) There are twelve dimensions of leader behavior assessed using nine individual and group exercises. In addition, nine psychometric tests are used as an integral part of assessment. Assessment occurs within a three-day total immersion evaluative situation.

The use of the assessment process for professional development was tested on two populations: students entering the Infantry Officer's Advanced Course (IOAC) and the sutdents entering the Advanced Non-Commissioned Officer's Educational System (ANCOES). Because these two populations have different professional experiences and organizational expectations, the assessment exercises developed were different (although parallel).

PROCEDURE

Operation: ACTR assessed 24 groups of assessees. Each group comprised (usually) eighteen men. Ten of the groups were counseled. This phase occupied FY 74 and the first half of FY 75.

Follow-Up: The students from BIOCC, IOBC, IOAC, and ANCOES who attend ACTR for evaluation and/or counseling and their classmates who have been designated as control subjects will be studied in a follow-up research program for two years after assessment.

Present Paper: The present paper addresses the question: Does the assessment/counseling program (for ANCOES and IOAC) result in a change in self-concept? Is assessment a learning process?

Both before entry into ACTR and after counseling at ACTR self-rating scales are administered in six areas: social, oral, writing, leadership, administrative, and problem solving skills.

Simply put, the question addressed in this preliminary analysis is, does the assessment/counseling process result in a shift on the six scales?

RESULTS

Self-Ratings: The data are strikingly clear. In both ANCOES and IOAC groups, the shift on each of the six scales is significant.

Relationships Between Self-Ratings and Assessor Ratings: The comparison between self-rating at entry to ACTR and assessor ratings indicates a significant relationship in one case for IOAC (Problem Solving) and in no case for ANCOES. The comparison between self-ratings after counseling and assessor ratings indicates significant relationships in three cases for IOAC assessees and four cases for ANCOES.

DISCUSSION

The shifts in self-ratings are offered as evidence of the affect that the assessment/counseling experience has on self-opinion. When the assessee enter ACTR, his self-rating scores, on the average, are higher than when he leaves ACTR. He has learned about himself.

The relationships between assessee self-ratings at entry (before the assessment program begins) and assessor ratings (composite scores representing the cumulative observations of the several assessors over the several exercises) indicate that only the ratings by and of the IOAC assessees on Problem Solving are in agreement (in terms of systematically varying directly proportionately with each other). After counseling, the agreement is better. A total of seven significant relationships is found. It is likely that the ACTR assessment/counseling program plays a role in effecting the change by the assessees towards better agreement with the assessors. It is probable that some shift in self-ratings results from the perception that the assessee has of himself as he performs in the assessment exercise. It is possible that being counseled also contributes to the change in self-perception.

FACTOR ANALYSIS OF LEADERSHIP MEASURES FROM A TRAINING PROGRAM

by

Ronald G. Downey Ph.D.

A Paper Prepared for Presentation at the
16th Annual Military Testing Association (MTA) Conference
October 21-25, 1974

Individual Training and Performance Research Laboratory
U.S. Army Research Institute for the Behavioral and Social Sciences
Arlington, Virginia 22209

FACTOR ANALYSIS OF LEADERSHIP MEASURES FROM A TRAINING PROGRAM

Ronald G. Downey Ph.D.

U.S. Army Research Institute for the Behavioral and Social Sciences
Arlington, Virginia 22209

Recent leadership research has focused upon the interactive influence of the leader and the organizational structure in which he operated (see Jacobs[1] for a review of these efforts), and results have indicated the difficulty of identifying the effective leader without considering the structure. On the other hand, research in the U.S. Army[2] has established broad behavioral dimensions of leadership in a simulated combat situation. The organizational requirements of the Army make essential the early and continuous identification of leaders who can perform effectively in a variety of situations. Current methods of evaluation depend heavily on job performance ratings, which are subject to a variety of problems. Expanded opportunities do exist in Army training programs for evaluating leadership potential; however, measures currently are used in an unorganized and unsystemized way.

This paper reports an effort to identify the implicit dimensions of leadership being evaluated by the training staff as part of an Army officer training program. This was accomplished by factor analysis of the measures being used by the school staff.

METHOD

SAMPLE

The sample population was 244 newly commissioned officers attending Officer Basic Course for Engineers. Training consisted of a nine-week course, teaching basic technical, managerial, and leadership knowledge and skills. Most officers were reserve 2d lieutenants on active duty only for the training period. They attended classes for approximately 8 hours a day 5 days a week.

---

[1] Jacobs, T. O. Leadership and exchange in formal organizations. Alexandria, Virginia: Human Resources Research Organization, 1971.

[2] Helme, W. H., L. P. Willemin, and F. C. Grafton. Dimensions of leadership in a simulated combat situation. ARI Technical Research Report 1172, 1971.

ANALYSIS

A principal components factor analysis was done with the highest off-diagonal correlation coefficient used as the communality estimate. A varimax rotation was then done on all factors having an eigen value greater than 1.00. Due to missing data, the correlation coefficients were based on slightly variable N's; the maximum N for each pair of variables was used for the analysis. The smallest N for any correlation was 240 and the largest N was 244. Rotation solutions were also looked at by reducing the number of original factors by one for factor solutions greater than two.

RESULTS

Table 1 gives the intercorrelation matrix for all variables, and Table 2 gives the rotated factor loading for the three factor solution adopted. Only loadings of an absolute value of .25 or greater are reported. The proportions of variance accounted for by the rotated factors are reported at the bottom of Table 2.

Factor I, called "cognitive/verbal leadership skills," has the highest loading on the cognitive portions of the OEB and the training paper-and-pencil tests. Factor II, called "leadership behavioral style," has the highest loadings on the associate nominations, TAC officer ratings, and the non-cognitive portions of the OEB; these measures reflect the manner in which leadership is displayed. Factor III, called "leadership performance," has the highest loading on the CEPE (T) and (L) field problems and the Engineer Stakes (L). The school tests loaded on both Factor I and Factor III; the associate nominations and TAC evaluations loaded on both Factor II and Factor III.

These three factors were closely related to evaluative concepts utilized in training and assessment programs.[3] Cognitive/verbal leadership skills was related to academic orientation or general mental abilities. Leadership behavioral style was related to evaluations of characteristics by self, peers, and superiors. Finally, leadership performance was related to applied exercises evaluating performance.

---

[3] Wollowick, H. B. and W. J. McNamara. Relationship of the components of an assessment center to management success. Journal of Applied Psychology, 1969, 53, 348-352.

Table 1

INTERCORRELATION COEFFICIENTS OF ALL VARIABLES

Variable Name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. OEB Combat Ldrs (C)	1.00																	
2. CEPE (L)	.15*	1.00																
3. OEB Combat Ldrs (N-C)	.13*	.12	1.00															
4. Combat Operations Exam	.30*	.21*	.01	1.00														
5. Orienteering	.11	.18*	.23*	.04	1.00													
6. OEB Career Potential (N-C)	.25*	-.08	.41*	-.05	.14*	1.00												
7. OEB Tech-Mgr Ldrs (C)	.37*	.12	.02	.25*	-.02	.04	1.00											
8. CEPE (T)	.21*	.74*	.12	.26*	.22*	-.06	.12	1.00										
9. OEB Career Potential (C)	.14*	.08	.11	.16*	.07	.18*	.42*	.10	1.00									
10. OEB Tech-Mgr Ldrs (N-C)	.02	.28*	.29*	.09	.07	.02	.25*	.23*	.22*	1.00								
11. Leadership Exam	.28*	.15*	.05	.50*	.03	-.07	.19*	.25*	.19*	.15*	1.00							
12. Horizontal Construction Exam	.36*	.36*	.04	.39*	.05	-.05	.29*	.44*	.29*	.11	.29*	1.00						
13. 8th Week Associate Nomination	.16*	.29*	.24*	.15*	.22*	.10	-.09	.23*	.01	.13	.23*	.13*	1.00					
14. TAC Officer Rating	.17*	.20*	.27*	.20*	.11	.17*	-.02	.20*	.13*	.23*	.15*	.14*	.59*	1.00				
15. OEB Career Intent	.07	.01	.35*	-.12	.11	.36*	-.10	-.02	.12	.11	-.00	-.17*	.19*	.22*	1.00			
16. Engineer Stakes (T)	.33*	.38*	.14*	.32*	.19*	.06	.20*	.39*	.32*	.16*	.24*	.37*	.19*	.20*	.04	1.00		
17. Engineer Stakes (L)	.14*	.36*	.15*	.21*	.19*	-.05	.02	.34*	.12	.18*	.20*	.22*	.30*	.10	.05	.20*	1.00	
18. Oral Presentation (History)	.19*	.17*	-.02	.25*	-.08	.08	.14*	.11	.12	.16*	.16*	.17*	.09	.18*	.07	.02	.01	1.00

*significant at P <.05

Table 2

FACTOR ANALYSIS OF VARIABLES

Variable Name	ROTATED FACTOR LOADINGS[a] I	II	III
OEB Combat Leadership (C)	.67		
CEPE(L)			.78
OEB Combat Leadership (N-C)		.61	
Combat Operations Exam	.51		.31
Orienteering		.26	.25
OEB Career Potential (N-C)		.58	
OEB Technical-Managerial Ldrs (C)	.61		
CEPE (T)			.78
OEB Career Potential (C)	.64		
OEB Technical-Managerial Ldrs (N-C)			.25
Leadership Exam	.43		.29
Horizontal Construction Exam	.52		.39
8th Week Associate Nomination		.47	.49
TAC Officer Rating		.52	.36
OEB Career Intent		.57	
Engineer Stakes (T)	.41		.38
Engineer Stakes (L)			.45
Oral Presentation (History)	.26		
Proportion of Variance	.13	.10	.14

---

[a] Only loadings of absolute value of ≤.25 were reported.

A two-factor solution resulted in a drop in the total variance accounted for by the solution, from 37% for three factors down to 28% for two factors, and resulted in the leadership behavioral style factor remaining intact while the cognitive/verbal leadership skills and leadership performance combined into one factor. There were only two variables with moderate loadings on both factors; all other variables loaded substantially on only one factor.

CONCLUSIONS

Several possible difficulties in this analysis should be pointed out. In general the measures used were not evaluated in terms of reliability, and the low level of intercorrelations would suggest that some measures did in fact suffer from rather low reliability. (First-hand observation of some procedures has supported this view.) Only the OEB[4] and the associate nominations[5] have reliability estimates (.70's to .90's). Preliminary work with the school had encountered some difficulties in evaluation time available and a generalized tendency for "halo" effects on evaluations. Finally, the constraints imposed by an operational setting precluded a more integrated and systematic approach to the inclusion of leadership measures.

Although this study is preliminary, three broad leadership factors were identified. A two factor solution resulted in loss of variance accounted for. Pending more definitive analysis, these factors were cognitive/verbal leadership skills, leadership behavioral style, and leadership performance. Further research is under way to verify this structure in other similar training programs, to improve the measurement qualities of the variables, and, finally, to study the relationship between the identified factors and performance in a variety of leadership positions.

---

[4] US Army Research Institute for the Behavioral and Social Sciences, Manual for interpreting the Officer Evaluation Battery. Arlington, Virginia, 1973.

[5] Downey, R. G. Associate evaluations: Nominations vs. ratings. ARI Technical Paper 253, 1974.

REFERENCES

Downey, R. G. Associate evaluations: Nominations vs. ratings, ARI Technical Paper 253. 1974.

Helme, W. H., L. P. Willemin, and F. C. Grafton, Dimensions of leadership in a simulated combat situation. ARI Technical Research Report 1172, 1971.

Jacobs, T. O. Leadership and exchange in formal organizations. Alexandria, Virginia: Human Resources Research Organization, 1971.

US Army Research Institute for the Behavioral and Social Sciences, Manual for interpreting the Officer Evaluation Battery, Arlington, Virginia, 1973.

Wollowick, H. B. and W. J. McNamara, Relationship of the components of an assessment center to management success. Journal of Applied Psychology, 1969, 53, 348-352.

AN EXPERIMENTAL ASSESSMENT PROCEDURE FOR EVALUATING OVERSEAS PERFORMANCE

Ted M. I. Yellen
Navy Personnel Research and Development Center
San Diego, California

This presentation will consist primarily of viewing a 15 minute video tape of an experimental assessment procedure for evaluating overseas performance. The tape documents the administration of a Japanese Contact and Communication Course.

Before we start with the tape I would like to take a few minutes and provide a brief background and overview of the overseas performance assessment project.

As part of a research program designed to develop a selection procedure for overseas assignment (Yellen & Mumford, 1974), the Navy Personnel Research and Development Center in conjunction with Simile II of La Jolla, California, developed the Japanese Contact and Communication Course. As originally conceived, the course was to serve as a criterion measure of overseas adjustment. However, as the simulation was being developed and tested with Navy personnel stationed in San Diego, it became apparent that it could also serve as an overseas training device.

As a criterion measure, the simulation gives Navy personnel stationed in Japan an experience in communicating with Japanese people in controlled situations as a means of observing their performance and obtaining their attitudes, reactions, and opinions toward their tour of duty. The procedure involves a number of situations which portray various aspects of the Japanese culture. These situations represent activities which Navy personnel should have some knowledge of and experience with in their interaction with Japanese people and in their every day living in Japan. The participants are given a limited amount of Japanese money to complete six simulated situations. For example, one of the situations involves a train station ticket counter in which the Navy person has the task of buying a ticket from a Japanese ticket agent to a specific city. The task not only requires the purchase of the ticket but being able to translate to another person the exact route and related train changes. The Japanese people at each of the six stations explain everything in detail, however, they speak only Japanese. Other situations are centered around the interaction of a Navy person with a Japanese waiter, hotel clerk, theater ticket seller, and shopkeeper. Also the Navy person is required to purchase a subway ticket in Tokyo. Upon completion of each task the individual reports to a "feedback table" where he is told if he successfully completed the task. The feedback table is manned by a Japanese person who can read Japanese and an assistant director who can speak the Japanese language and has a good understanding of their culture and customs. They are responsible for evaluating the performance of the individuals participating in the simulation exercise.

The assessment procedure for evaluating overseas performance involved the development of two separate measurement techniques. The first method consists of evaluating the individual's performance on each of the six tasks using a three point rating scale. For example, upon completion of the train ticket exercise the Japanese ticket agent writes in Japanese the price, route, and destination on a receipt and gives it to the participant. He then goes to the feedback table, gives the ticket to the individuals at the table and explains to them the price, route, and destination he thinks he has taken. If there is no discrepancy between what he thinks he has done and what is actually on the receipt, then the evaluators punch a "2" on his ticket. If he accomplished the task, but did not understand fully what he was doing, then he receives a "1" on his ticket. If he had no idea of what he was doing and did not accomplish the task, he receives a zero. The evaluators keep a record of his score and also provide him with immediate feedback as to his ability to negotiate the transaction.

The second measurement technique is a self-evaluation instrument entitled the "Navy Overseas Adjustment Scale" (NOAS). This scale is designed to provide information on a participant's level of effective functioning in the Japanese culture. More specifically, the measurement procedure obtains evaluative-type information concerning the participant's attitude toward Japanese people, the amount of personal interaction with Japanese people, their feelings concerning the Japanese culture, and, in general, their feelings about living in Japan. The development of the adjustment scale followed basically the mixed standard scale procedure. As described by Blanz and Ghiselli (1972), this method is designed to minimize the errors in rating and provides an index of the accuracy of rating. Briefly stated, there are three descriptions representing degrees of overseas adaptation on ten important traits. The individual responds by indicating the extent to which each description fits him, over-estimates his adaptation or under-estimates his adaptation.

It has been our experience, although not verified experimentally, that participants are more willing to report their experiences more honestly and in greater detail after participating in the simulation than when asked to report their feelings about Japan without having participated in the simulation experience.

The validity of the assessment procedure will be determined by correlating the performance evaluations and NOAS scores with command and peer nominations. The command nomination procedure involves an overseas command in Japan providing a list of individuals who have come to the command's attention for detrimental reasons. Also, the command will provide the names of individuals who are doing an outstanding job in Japanese-American community relations affairs. The peer nomination procedure involves having personnel nominating individuals who they feel have adapted extremely well to living in Japan and also listing those that they feel are having an extremely difficult time adjusting. Individuals nominated as outstanding and those having adjustment problems, as well as individuals in between, will be requested to participate in the Japanese Contact and Communication Course exercise. These individuals will serve as the high and low criterion groups for the validation effort.

As previously mentioned, the simulation was designed as a criterion measure of overseas adjustment. However, in the developmental phases of the simulation, the course has shown successful application as a training vehicle to introduce the Japanese culture to personnel who are preparing to go to Japan. When the course is used in this manner, then, upon completion of the tasks a debriefing is conducted. The discussion is aimed at helping the participants understand their attitudes, manners, and demeanor which facilitate communication with the Japanese and those which create problems. These discussions are tailored to the needs of the participants.

REFERENCES

Blanz, F. A. and Ghiselli, E. E. The Mixed Standard Scale: A New Rating System. Personnel Psychology, 1972, 25, 185-199.

Yellen. T. M. I. and Mumford, S. J. The Cross-Cultural Interaction Inventory: Development of Overseas Criterion Measures and Items that Differentiate Between Successful and Unsuccessful Adjusters. Navy Personnel Research and Development Center, San Diego, California (to be released in the Fall of 1974)

SECTION XIII

MANAGEMENT AND MEASUREMENT PROBLEMS

RESPONSE BIAS
IN
WORKER TASK INVENTORIES

by

Major W.E. Slack, USMC

During 1972 a casual observation was made that Gunnery Sergeants (paygrade E7) performed many tasks and spent "very much" of their time on each task performed. The observation was made while viewing the printout of raw task inventory data input to the Computerized Occupational Data Analysis Programs (CODAP). While monitoring subsequent studies the premise appeared to hold true, therefore it was decided that this "Gunny Syndrome" theory should be looked more closely to ascertain whether this was in fact the case and, if so, was it common to all fields.

Computer programs were written to extract raw case data from the original raw data file for the study and subsequently develop statistical summaries of response values (1-7) and numbers of responses. The extraction program was developed such that the response value of any byte position in the record could be used as the extraction key by making Boolean comparisons. This afforded the opportunity to extract and summarize data based on any question answered by the individual.

In this first look at response bias, data was summarized by paygrade in four occupational studies. One study was further briken into known and anonymous respondee categories by paygrade. The four occupational fields selected for analysis were studied after early 1972. Data from most earlier studies were deleted from the data bank upon approval of the study report and after action had commenced to implement the recommendations set forth therein. A change in office policy, which provided for the retention of the major files of an occupational field study, has eliminated this problem of data availability for its use in analyses of this nature. In addition, it should give us some benchmarks for the next study cycle when the field is studied again.

At the outset let me make it clear that I am not a statistician nor am I knowledgeable in the areas of psychology or educational testing and evaluation. The idea behind this paper is to provide some food for thought which could result in answers to questions I have, in addition to possibly enhance task inventory development and administration.

The four occupational fields selected were:

OF33 - Food Services
OF40 - Data Systems
OF46 - Photography
257X - Special Communications.

The Special Communications group was originally a series of Military Occupational Specialties (MOS's) within the Communications Occupational Field (OF25) but has since been assigned its own occupational field designator (OF26). As can be seen there is a considerable difference, at least theoretically, between the fields investigated.

After some initial surprises as to the obvious distribution of responses (both values and numbers) I decided that, rather than attempting any great breakthroughs, I would run some basic tests, draw a few graphs, display a few tables of data and go back to the drawing boards to look at some different aspects which might cause the variability noted.

Tables 1 through 5 display the recap statistics for each occupational field by paygrade and include sample size, range of mean response values, sample mean response value, range of numbers of responses and sample mean number of responses. Tables 4 and 5 display the data for OF46 broken into two categories, anonymous and identifiable respondees. Tables 6 through 9 show the frequency distribution of response values by grade for each study, while Figures 9-12 ref;ect this data in graphic form. Tables 10 through 13 are included to indicate the manner in which the numbers of responses data were grouped and the result of that grouping. Figures 5 through 8 give a quick visual display of the resultant frequency distribution. Figures through 4 show the frequency distributions of grouped data on mean response values of individuals in the population.

The initial test conducted on the data was an analysis of variance (ANOVAR) F-test at the 0.05 significance level. The test was conducted on two factors of each study -

1. Response values and
2. Numbers of responses.

The null hypothesis in each test was:

> The means of the subpopulations of each field were not significantly different from the mean of the total field population.

In every occupational field the result was the same - accept the null hypothesis regarding mean response value and reject the null hypothesis when the mean number of responses distribution was considered.

In an effort to look more closely at the relationship of each subpopulation mean value to the occupational field's total population mean value, a series of t-tests were conducted. The null hypothesis in each test was:

> Each subpopulation (paygrade) mean was not significantly different from the population (occupational field) mean.

These tests were again conducted both on the mean response value means and on the mean numbers of responses. The latter more from curiosity than for any other reason. The findings bear out the results of the previous test with one "glitch", the rejection of the E5 values shown in Table 16. Table 14 through 21 show the results of this series of t-tests.

Given the results of these tests I can only conclude that there is no "Gunny Syndrome" as originally proposed. There appears to be some significance in the variations of the number of responses which will receive some additional attention in the future to determine its cause. However, I would not like to go out on a limb with any rash prognostications at this point of the game.

For the devotees of the "many point" scales of response values I would like to call attention to the distributions of response values shown in Figures 9 through 12. After viewing these graphs I must ask the question -- "What happens when you insert still finer gradations between the extreme points"?

Finally, I had hoped to include some thoughts and statistics on the question of anonymous versus identifiable respondees to task inventories. I did not feel that the data presently available would be adequate to broach the problem. We will be conducting a deeper look into this question withnew studies now underway wherein more care was taken in the administration of the questionnaire to ensure better samples across the board.

GRADE	N	RANGE OF INDIVIDUAL MEAN RESPONSE VALUES (IMRV)	MEAN RESPONSE VALUE	STANDARD DEVIATION OF IMRV	RANGE OF NUMBERS OF RESPONSES	MEAN NUMBER OF RESPONSES
E2	17	1.5849 - 6.3125	3.5290	2.1219	17 -135	48.71
E3	45	1.2597 - 5.8571	3.7204	2.0226	21 - 100	52.22
E4	150	1.2651 - 6.1111	3.4726	2.0110	13 - 173	49.41
E5	169	1.5870 - 6.7714	3.8644	2.0159	12 - 142	51.57
E6	120	1.0909 - 6.0556	3.4434	1.9521	12 - 239	63.41
E7	85	1.4321 - 6.1340	3.4999	2.0190	12 - 160	66.27
E8	38	1.4805 - 6.8889	3.4269	1.9435	9 - 128	46.33
E9	12	2.6462 - 4.2989	3.5827	1.8206	14 - 87	46.33

TABLE 1. Primary statistics for 257X study.

GRADE	N	RANGE OF INDIVIDUAL MEAN RESPONSE VALUES (IMRV)	MEAN RESPONSE VALUE	STANDARD DEVIATION OF IMRV	RANGE OF NUMBERS OF RESPONSES	MEAN NUMBER OF RESPONSES
E1	115	1.0000 - 7.0000	4.0493	1.7001	1 - 84	22.58
E2	404	1.0000 - 7.0000	4.1847	1.8141	1 - 107	25.59
E3	328	1.0000 - 7.0000	4.1709	1.8980	1 - 305	31.30
E4	174	1.7500 - 7.0000	4.3339	1.7966	1 - 273	29.11
E5	235	2.4000 - 7.0000	4.6126	1.8853	2 - 131	30.47
E6	168	1.7288 - 7.0000	4.5129	1.9205	1 - 150	39.35
E7	64	1.9107 - 6.8224	4.3125	2.0384	3 - 125	42.09
E8	20	2.3412 - 6.2000	4.1947	1.7851	17 - 87	46.75
E9	13	1.9091 - 5.6111	3.8917	1.7341	14 - 59	39.08

TABLE 2. Primary statistics for OF33 study.

GRADE	N	RANGE OF INDIVIDUAL MEAN RESPONSE VALUES (IMRV)	MEAN RESPONSE VALUE	STANDARD DEVIATION OF IMRV	RANGE OF NUMBERS OF RESPONSES	MEAN NUMBER OF RESPONSES
E1	20	1.0000 - 6.4286	3.5444	2.0349	3 - 57	24.80
E2	139	1.0417 - 6.8261	3.7296	1.9694	6 - 102	33.26
E3	226	1.0000 - 6.8462	3.7590	1.9541	8 - 131	40.61
E4	135	1.0000 - 6.1200	3.4816	1.9342	5 - 206	52.17
E5	163	1.0000 - 7.0000	3.6705	1.9689	6 - 193	56.04
E6	97	1.2500 - 6.5917	3.7090	1.9597	12 - 174	62.29
E7	77	1.5698 - 6.9605	3.5496	1.8969	16 - 212	64.96
E8	50	2.0000 - 5.3500	3.1452	1.8290	21 - 188	64.02
E9	18	1.5325 - 4.7983	3.2399	1.7942	27 - 119	64.83

TABLE 3. Primary statistics for 0F40 study.

GRADE	N	RANGE OF INDIVIDUAL MEAN RESPONSE VALUES (IMRV)	MEAN RESPONSE VALUE	STANDARD DEVIATION OF IMRV	RANGE OF NUMBERS OF RESPONSES	MEAN NUMBER OF RESPONSES
E1	-	-	-	-	-	-
E2	1	**N/A**	5.2254	0	N/A	71
E3	5	2.1429 - 5.3205	3.8527	1.8471	23 - 100	63.80
E4	16	2.7500 - 5.3409	3.6460	1.7773	19 - 78	49.44
E5	7	2.1913 - 5.8125	3.9901	2.0633	36 - 140	72.43
E6	13	2.5986 - 6.0495	3.9961	1.8300	28 - 171	77.92
E7	4	2.6190 - 6.2656	4.5401	2.0892	63 - 192	102.75
E8	3	3.1000 - 4.1429	3.6188	1.6555	42 - 90	60.33
E9	2	4.6528 - 5.1111	4.8056	2.4697	36 - 72	54.00

TABLE 4. Primary statistics for OF46 study (anonymous).

GRADE	N	RANGE OF INDIVIDUAL MEAN RESPONSE VALUES (IMRV)	MEAN RESPONSE VALUE	STANDARD DEVIATION OF IMRV	RANGE OF NUMBERS OF RESPONSES	MEAN NUMBER OF RESPONSES
E1	1	N/A	4.8000	N/A	N/A	10
E2	15	1.3226 - 4.3871	3.6380	1.5875	6 - 90	33.33
E3	16	1.8684 - 5.3879	3.9678	1.8555	10 - 116	44.63
E4	29	1.6763 - 6.2222	3.7628	2.0797	15 - 139	54.66
E5	38	1.2921 - 6.0870	3.4598	1.9156	11 - 135	68.16
E6	36	1.3846 - 5.6000	3.8533	1.9573	9 - 194	76.86
E7	10	1.7045 - 4.1270	3.2809	1.6376	26 - 151	79.40
E8	17	1.6897 - 5.4000	3.6142	1.9175	20 - 201	81.88
E9	4	1.5510 - 3.8750	2.8371	1.9062	48 - 103	78.25

TABLE 5. Primary statistics for OF46 study (identifiable).

GRADE	RESP =1	RESP =2	RESP =3	RESP =4	RESP =5	RESP =6	RESP =7
E2	246	66	60	236	39	42	139
E3	573	171	132	824	123	168	359
E4	2033	771	551	2196	435	430	996
E5	1757	830	600	2812	556	687	1473
E6	1929	949	723	2141	448	532	887
E7	1510	638	417	1545	324	515	684
E8	611	339	212	738	118	163	288
E9	103	80	49	193	28	51	52

TABLE 6. Distribution of responses by grade(257X).

GRADE	RESP =1	RESP =2	RESP =3	RESP =4	RESP =5	RESP =6	RESP =7
E1	356	126	199	1201	238	142	355
E2	1257	671	776	4204	904	689	1839
E3	1491	612	773	3773	934	759	1924
E4	533	291	342	2014	546	349	990
E5	721	316	358	2636	600	650	1880
E6	774	311	292	2474	590	471	1698
E7	438	140	151	890	209	220	646
E8	133	36	35	411	88	96	136
E9	76	41	30	228	37	40	56

TABLE 7. Distribution of responses by grade (OF33).

GRADE	RESP =1	RESP =2	RESP =3	RESP =4	RESP =5	RESP =6	RESP =7
E1	140	40	20	171	27	27	71
E2	1014	405	350	1571	288	321	674
E3	1890	907	702	3022	657	680	1319
E4	1735	819	681	2049	461	482	816
E5	1949	1023	952	2626	622	673	1289
E6	1284	651	493	1860	426	521	807
E7	1183	463	467	1661	291	413	524
E8	829	296	204	1206	205	177	284
E9	291	174	109	399	40	59	95

TABLE 8. Distribution of responses by grade (OF40).

GRADE	RESP =1	RESP =2	RESP =3	RESP =4	RESP =5	RESP =6	RESP =7
E1	2	-	-	3	-	1	4
E2	84	47	40	258	50	34	58
E3	134	125	93	404	50	56	171
E4	492	259	267	635	182	188	351
E5	748	341	259	948	213	196	392
E6	625	389	380	1248	269	253	616
E7	228	123	147	389	80	66	172
E8	310	227	171	402	141	158	164
E9	150	38	25	106	7	18	77

TABLE 9. Distribution of responses by grade (OF46-all).

GRADE / NRESP	E1	E2	E3	E4	E5	E6	E7	E8	E9	TOTAL
1 - 10								1		1
11 - 20		1		4	10	9	6	1	1	32
21 - 30		3	5	21	21	11	8	2	3	74
31 - 40		2	13	35	35	16	9		1	111
41 - 50		6	6	32	33	18	12	5	2	114
51 - 60		2	7	18	23	14	8	8	1	81
61 - 70		1	5	17	13	18	5	8	2	69
71 - 80			4	12	15	5	9	4	1	50
81 - 90		1	2	6	3	6	8	2	1	29
91 - 100			3	3	7	7	4	2		26
101 - 110				1	2	1	4	3		11
111 - 120					2	2	4	1		9
121 - 130					3	4	4	1		12
131 - 140		1			1	1	1			4
141 - 150					1	2				3
151 - 160						3	3			6
161 - 170										-
171 -180				1		1				2
181 - 190						1				1
191 - 200										-
201 - 210										
211 -220										
221 - 230										
231 - 240						1				1
										636

TABLE 10. Grouped distribution of numbers of responses by grade (257X).

GRADE NRESP	E1	E2	E3	E4	E5	E6	E7	E8	E9	TOTAL
1-10	28	87	51	36	36	21	6			265
11-20	36	100	80	39	67	37	9	1	2	371
21-30	22	85	73	32	44	28	14	4	2	304
31-40	16	50	39	25	25	25	11	3	2	196
41-50	3	40	31	21	21	15	3	6	3	143
51-60	6	27	29	9	13	7	7	2	4	104
61-70	2	6	10	7	16	7	4	1		53
71-80	1	6	7		6	9	3			32
81-90	1	2	1		2	6	1	3		16
91-100			2	1	1	3				7
101-110		1		2	1	3	3			10
111-120			1	1			2			4
121-130					2	1	1			4
131-140			1		1	1				3
141-150						5				5
151-160										
161-170			1							1
171-180										
181-190										
191-200										
201-210										
211-220										
221-230										
231-240										
241-250										
251-260										
261-270										
271-280			1	1						2
281-290										
291-300										
301-310			1							1
										1521

TABLE 11. Grouped distribution of numbers of responses by grade (OF33).

GRADE / NRESP	E1	E2	E3	E4	E5	E6	E7	E8	E9	TOTAL
1-10	4	10	5	8	2					29
11-20	7	26	32	12	10	5	4			96
21-30	3	35	47	15	17	8	9	3	3	140
31-40	2	29	46	19	24	13	7	15	1	156
41-50	2	17	39	23	19	14	12	6		132
51-60	2	12	21	15	30	20	14	5	3	122
61-70		2	16	15	20	10	7	1	4	75
71-80		4	5	10	13	3	6	4	3	48
81-90		2	4		10	11	3	7	1	38
91-100		1	8	7	5	2	4	3	1	31
101-110		1	2	4	7	1	2	3	1	21
111-120				1	3	1	2			8
121-130				2	1	1	1			5
131-140			1	1	1	3	1	1		8
141-150				1		1		1		3
151-160						2	1			3
161-170				1		1	1			3
171-180						1	2			3
181-190								1		1
191-200					1					1
201-210				1						1
211-220							1			1
221-230										
231-240										
241-250										
251-260										
										925

TABLE 12. Grouped distribution of numbers of responses by grade (OF40).

GRADE NRESP	E1	E2	E3	E4	E5	E6	E7	E8	E9	TOTAL
1-10	1	3	1			2				7
11-20			1	3	2	1		1		8
21-30		4	5	6	2	5	1	1		24
31-40		4	3	9	7	6	3	2		34
41-50		1	1	8	4	4	2	2		22
51-60		2	5	4	6	7	1			25
61-70			1	4	4	1	3	3	1	17
71-80		1	1	6	4	2	2		1	17
81-90		1			3	3	1	5		13
91-100			1	3	5	1			1	11
101-110			1		2	4		2	1	10
110-120			1		2	3				6
121-130				1	2	3	2	1		9
131-140				1	2					3
141-150						2				2
151-160						1	1	1		3
161-170						1				1
171-180						2				2
181-190										
191-200						1	1			2
201-210								1		1
211-220										
221-230										
231-240										
										217

TABLE 13. Grouped distribution of numbers of responses by grade (OF46).

GRADE	N_i	M_i	σ^2	t	H_0
E1					
E2	17	3.5290	2.1219	0.1160	Accept
E3	45	3.7204	2.0226	0.41	Accept
E4	150	3.4726	2.0110	0.648	Accept
E5	169	3.8644	2.0159	1.5636	Accept
E6	120	3.4434	1.9521	0.7546	Accept
E7	85	3.4999	2.0190	0.3898	Accept
E8	38	3.4269	1.9435	0.4990	Accept
E9	12	3.5827	1.8206	0.0150	Accept

TABLE 14. Response value (257X). M=3.5912 6 =4.0268

GRADE	N_i	M_i	σ^2	t	H_0
E1					
E2	17	48.71	754.09	1.036	Accept
E3	45	42.22	479.91	1.055	Accept
E4	150	49.41	485.95	3.001	Reject
E5	169	51.57	651.33	1.888	Accept
E6	120	63.41	1616.43	1.929	Accept
E7	85	66.27	1325.58	2.490	Reject
E8	38	64.97	720.13	1.976	Accept
E9	12	46.33	522.61	1.3719	Reject

TABLE 15. Number of Responses (257X). M=55.93 6 =925.74

GRADE	N_i	M_i	σ^2	t	H
E1	115	4.0493	2.8903	0.8172	Accept
E2	404	4.1847	3.2910	1.0743	Accept
E3	328	4.1705	3.6024	1.0720	Accept
E4	174	4.3339	3.2278	0.2723	Accept
E5	235	4.6126	3.5544	2.4067	Reject
E6	168	4.5129	3.6883	1.3990	Accept
E7	64	4.3125	4.1551	0.0689	Accept
E8	20	4.1947	3.1866	0.2420	Accept
E9	13	3.8917	3.0071	0.8010	Accept

TABLE 16. Response Value (OF33). M=4.2945 6 =3.4714

GRADE	N_i	M_i	σ^2	t	H
E1	115	22.58	272.58	4.633	Reject
E2	404	25.55	317.55	4.325	Reject
E3	328	31.30	576.75	0.533	Accept
E4	174	29.11	765.08	0.568	Accept
E5	235	30.47	500.42	0.069	Accept
E6	168	39.35	1060.80	3.453	Reject
E7	64	42.09	894.61	3.067	Reject
E8	20	46.75	408.85	3.499	Reject
E9	13	39.08	208.22	2.068	Reject

TABLE 17. Number of responses (OF33). M=30.36 6 =650.72

GRADE	N_i	M_i	σ^2	t	H
E1	20	3.5444	4.1480	0.2076	Accept
E2	139	3.7296	3.8785	0.4871	Accept
E3	226	3.7590	3.8185	0.8041	Accept
E4	135	3.4816	3.7411	0.8978	Accept
E5	163	3.6705	3.8766	0.1691	Accept
E6	97	3.7090	3.8404	0.3181	Accept
E7	77	3.5496	3.5982	0.4083	Accept
E8	50	3.4152	3.3452	0.8438	Accept
E9	18	3.2399	3.2192	0.9146	Accept

TABLE 18. Response Value (OF40). M=3.6422 6 =3.7959

GRADE	N_i	M_i	σ^2	t	H
E1	20	24.80	252.81	6.553	Reject
E2	139	33.26	355.32	8.613	Reject
E3	226	40.61	484.44	5.046	Reject
E4	135	52.17	1111.56	0.838	Accept
E5	163	56.04	804.25	2.627	Reject
E6	97	62.29	1208.95	3.436	Reject
E7	77	64.96	1607.21	3.2596	Reject
E8	50	64.02	1149.21	2.912	Reject
E9	18	64.83	649.74	2.429	Reject

TABLE 19. Number of responses (OF40) M=49.61 6 =950.22

GRADE	N_i	M_i	σ^2	t	H
E1					
E2	16	3.7372	2.6654	0.034	Accept
E3	21	3.9404	3.4379	0.502	Accept
E4	45	3.7213	3.9136	0.003	Accept
E5	45	3.5423	3.7979	0.559	Accept
E6	49	3.8912	3.7070	0.550	Accept
E7	14	3.6407	3.4862	0.153	Accept
E8	20	3.6149	3.5364	0.238	Accept
E9	6	3.4933	5.3166	0.220	Accept

TABLE 20. Response Value (OF46). M=3.7222 6 +3.7011

GRADE	N_i	M_i	σ^2	t	H
E1					
E2	16	35.69	505.80	4.592	Reject
E3	21	45.19	823.69	2.282	Reject
E4	45	52.80	536.04	2.507	Reject
E5	45	68.82	1054.30	0.672	Accept
E6	49	77.14	2189.30	1.661	Accept
E7	14	86.07	2528.08	1.478	Accept
E8	20	78.65	1954.52	1.294	Accept
E9	6	70.17	625.00	0.444	Accept

TABLE 21. Number of responses (OF46). M=65.07 6 =1552.32

	257X	OF33	OF40	OF46	OF46 ANON	OF46 KNOWN
NTASK	687	323	619	412	412	412
E1		7.0	4.0	2.4		2.4
E2	7.1	7.9	5.4	8.7	17.2	8.1
E3	7.6	9.7	6.6	11.9	15.5	10.8
E4	7.2	9.0	8.4	12.8	12.0	13.3
E5	7.5	9.4	9.4	16.7	17.6	16.5
E6	9.2	12.2	10.1	18.7	18.9	18.7
E7	9.6	13.0	10.5	20.9	24.9	19.3
E8	9.5	14.5	10.3	19.1	14.6	19.9
E9	6.7	12.1	10.5	17.0	13.1	19.0

TABLE 22. Average Number of Responses as percent of Total Tasks in Inventory.

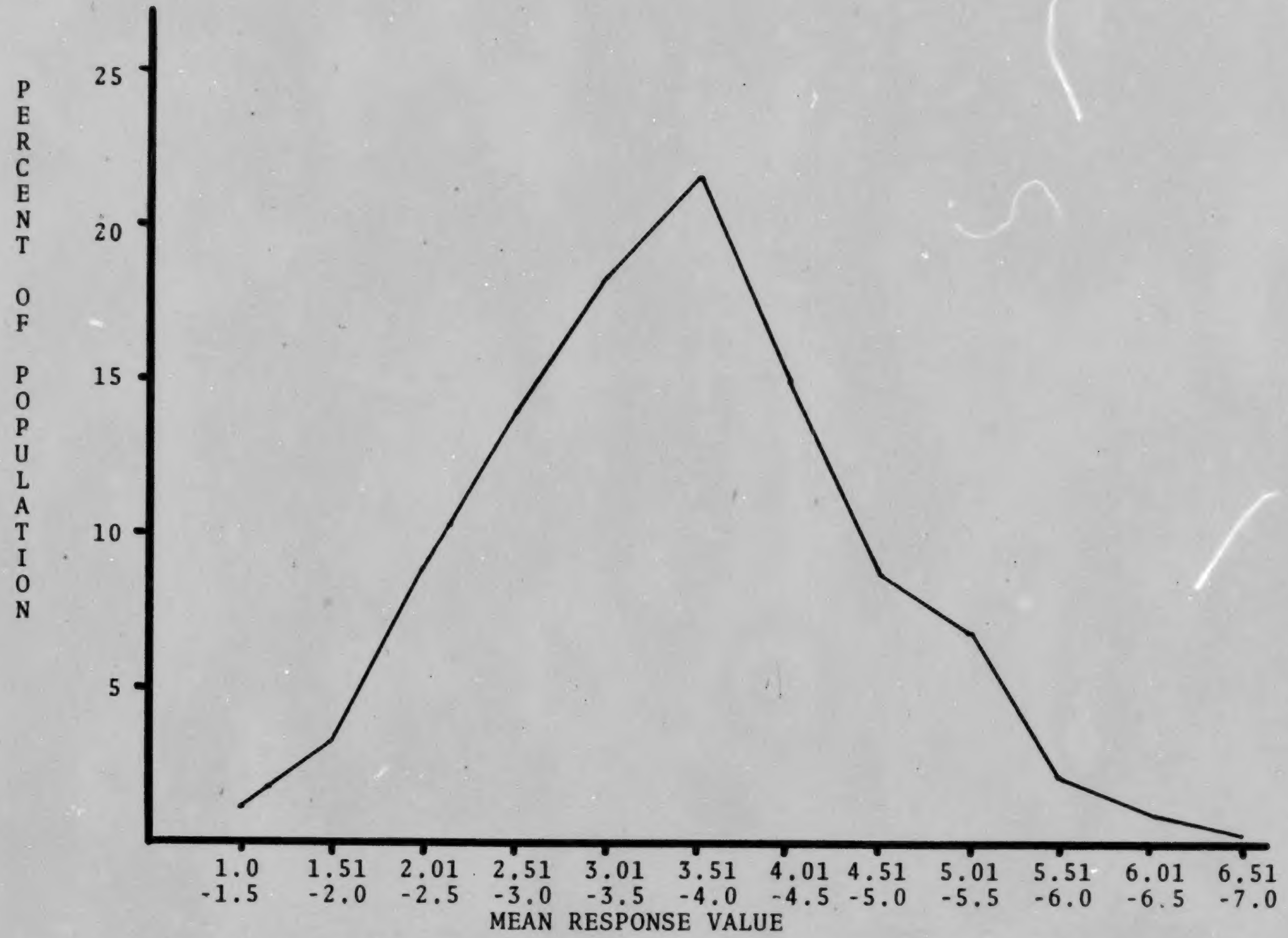


FIGURE 1. Distribution of mean response values (257X).

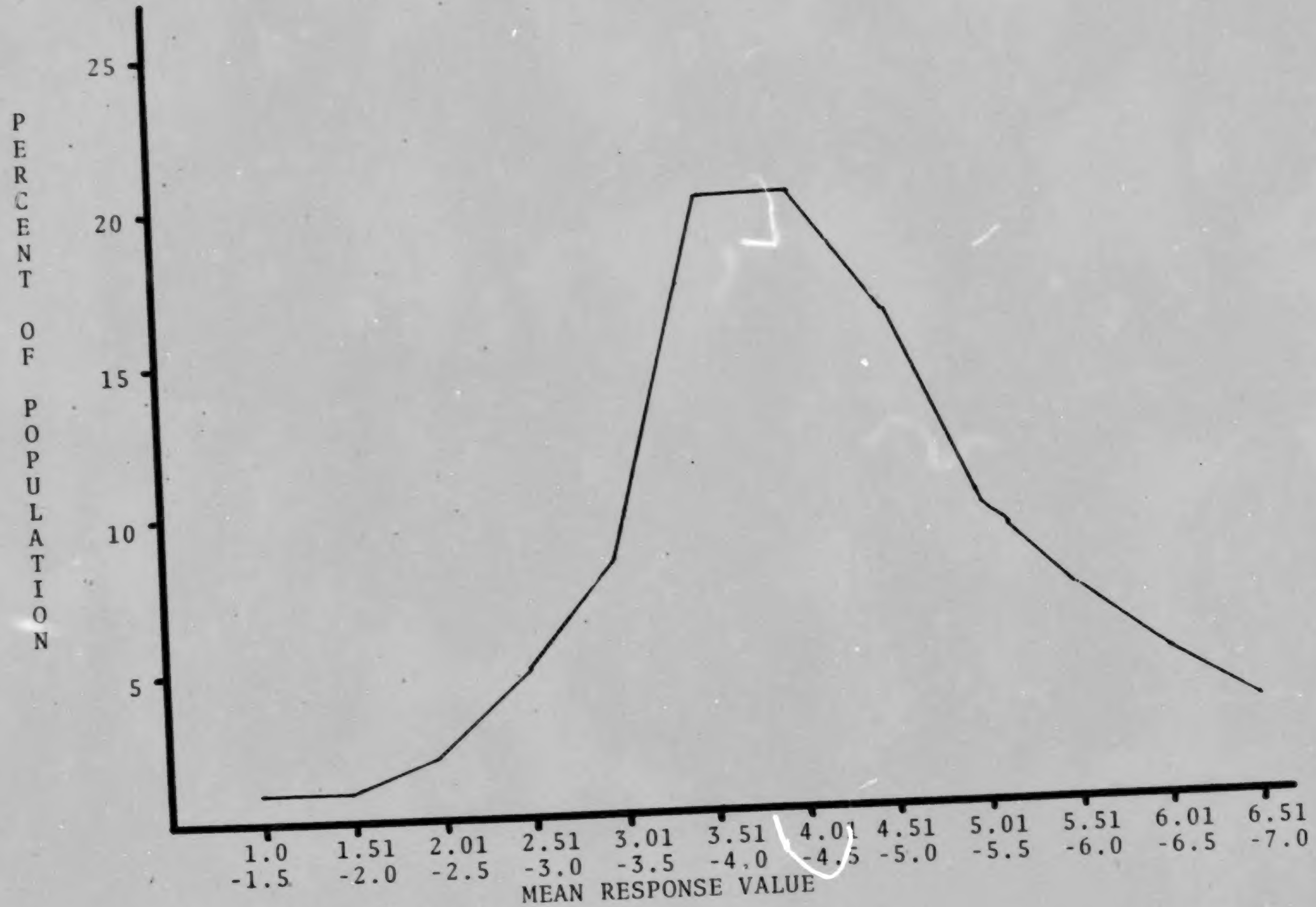


FIGURE 2. Distribution of mean response values (OF33).

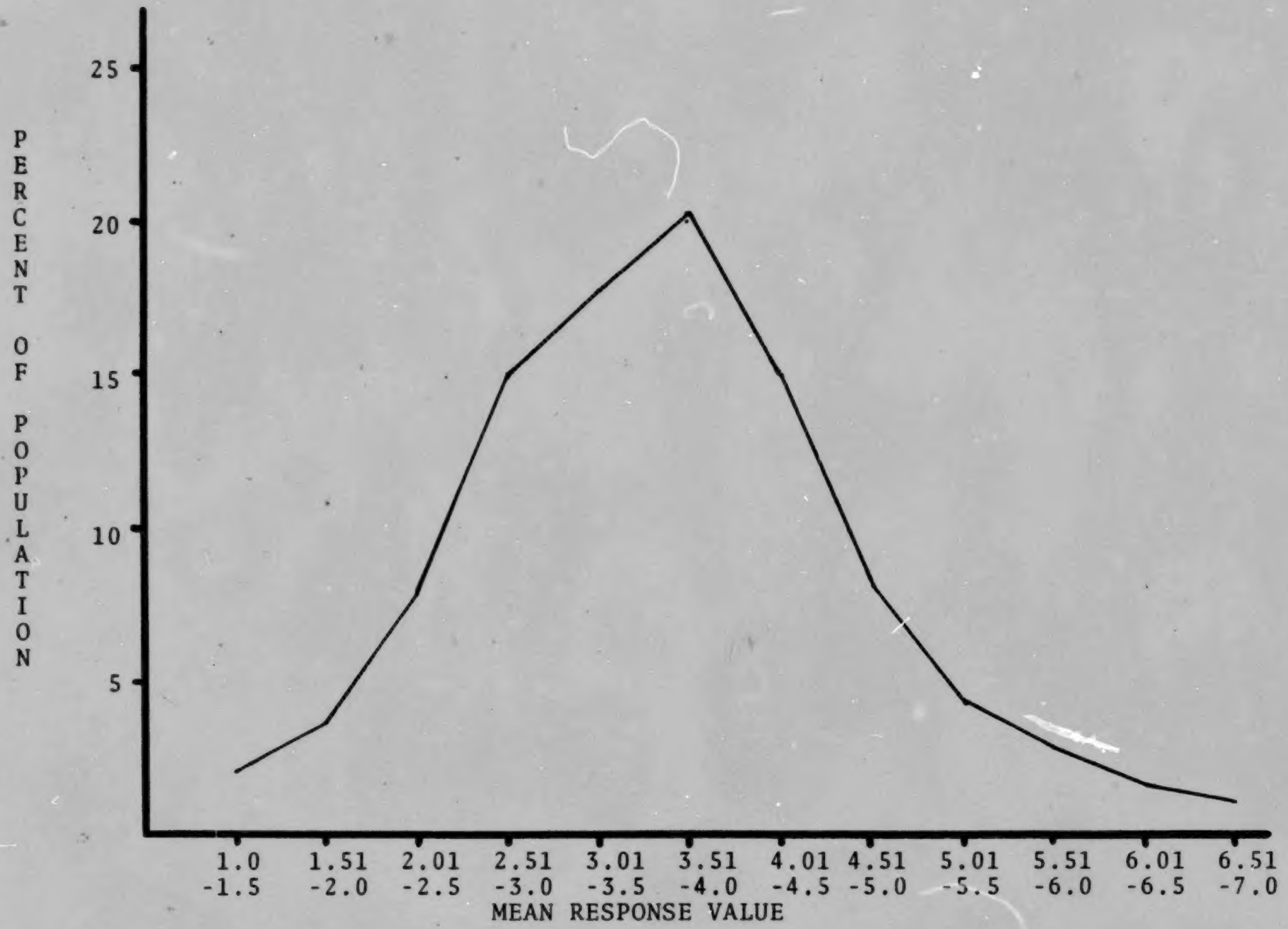


FIGURE 3. Distribution of mean response values (OF40).

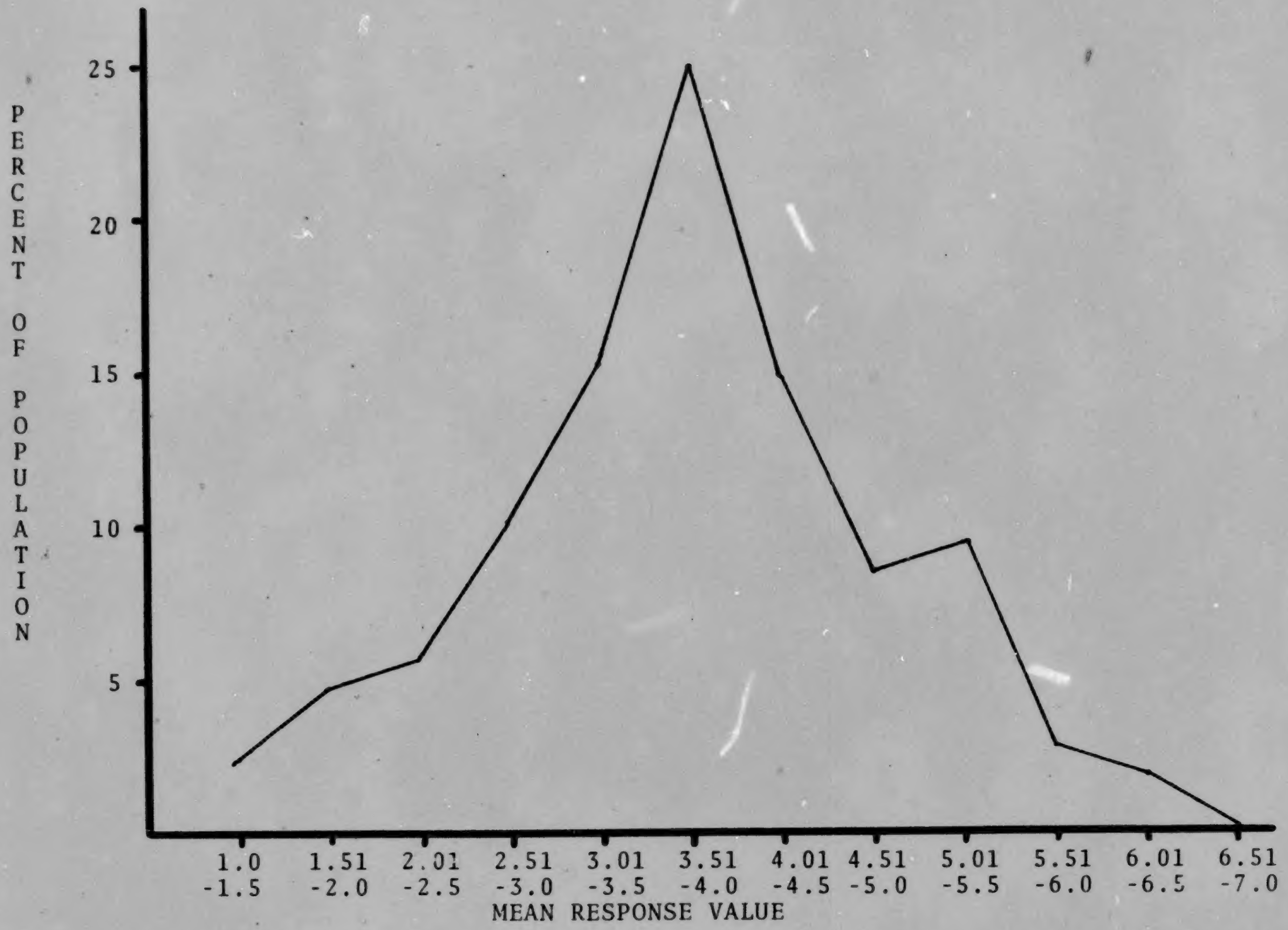


FIGURE 4. Distribution of mean response values (OF46).

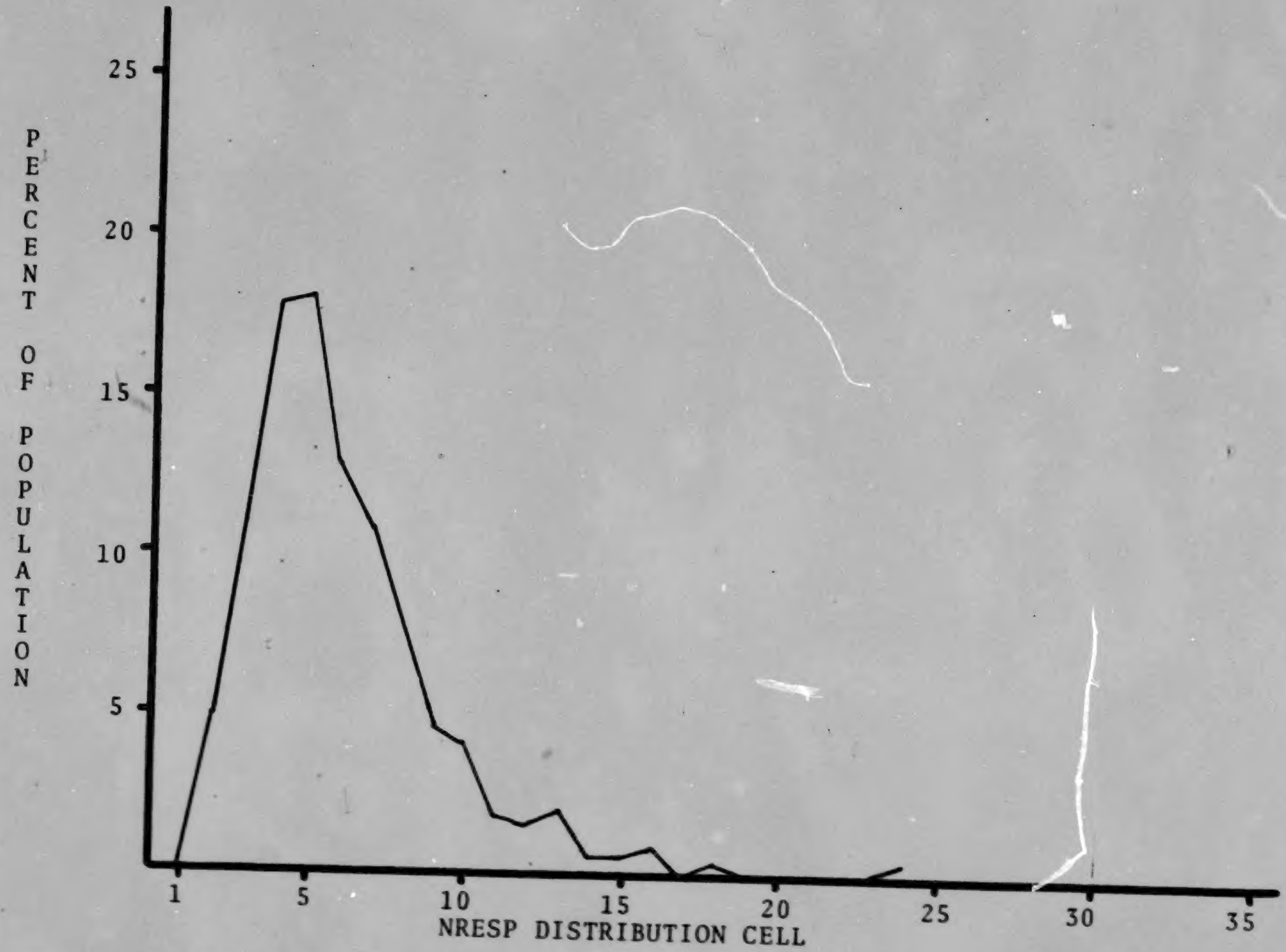


FIGURE 5. Distribution of numbers of responses (257X).

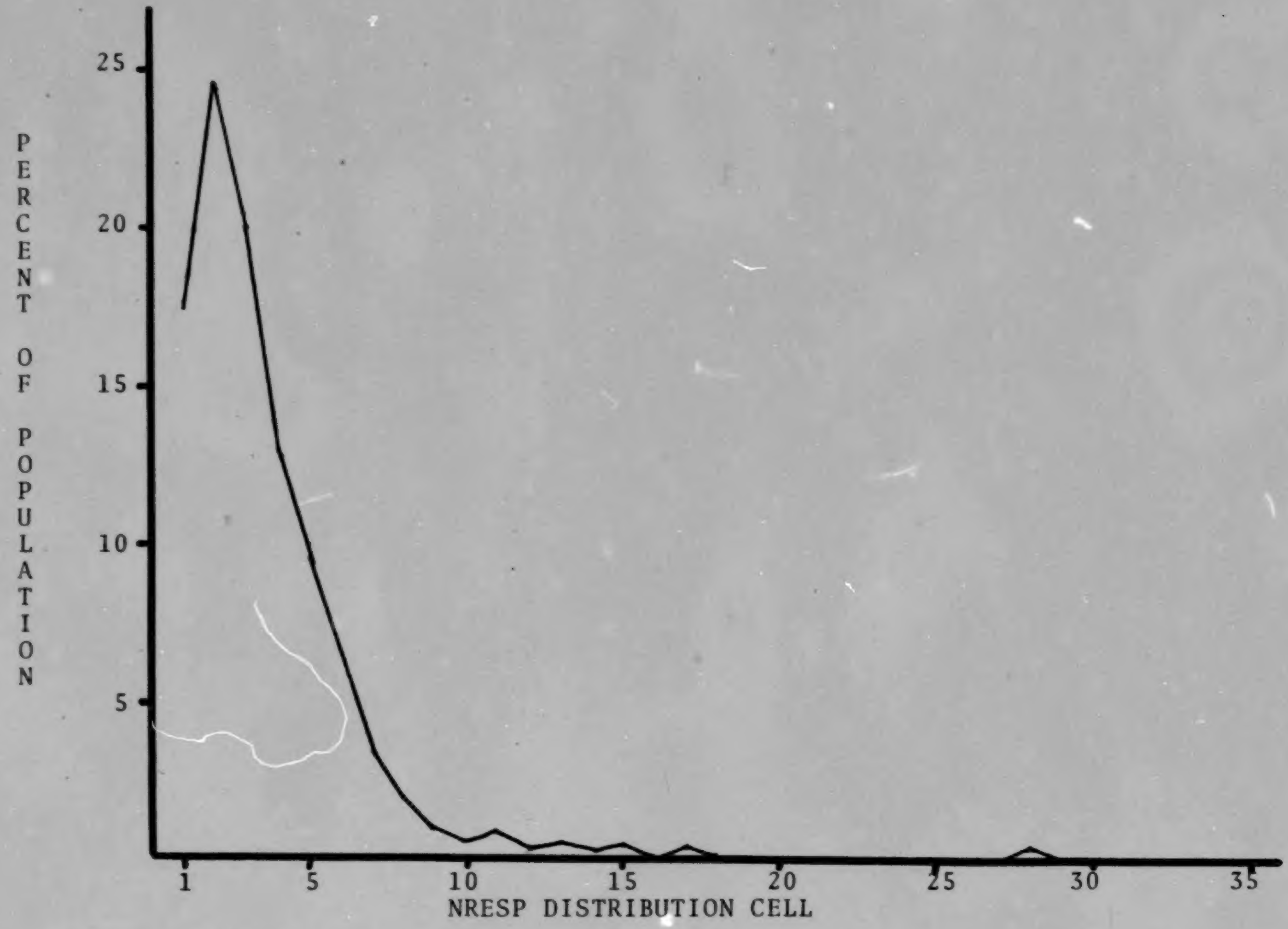


FIGURE 6. Distribution of numbers of responses (OF33).

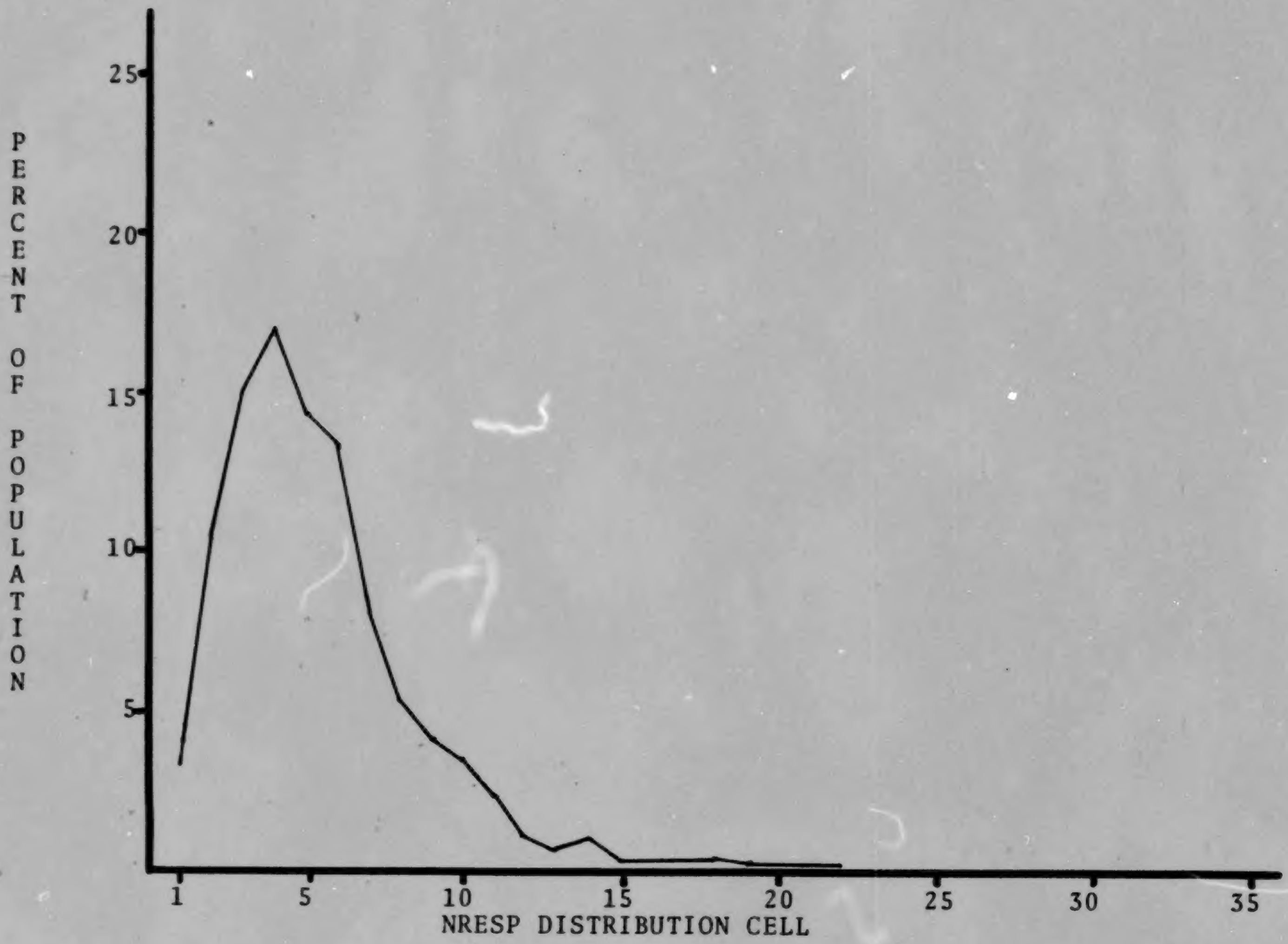


FIGURE 7. Distribution of numbers of responses (OF40).

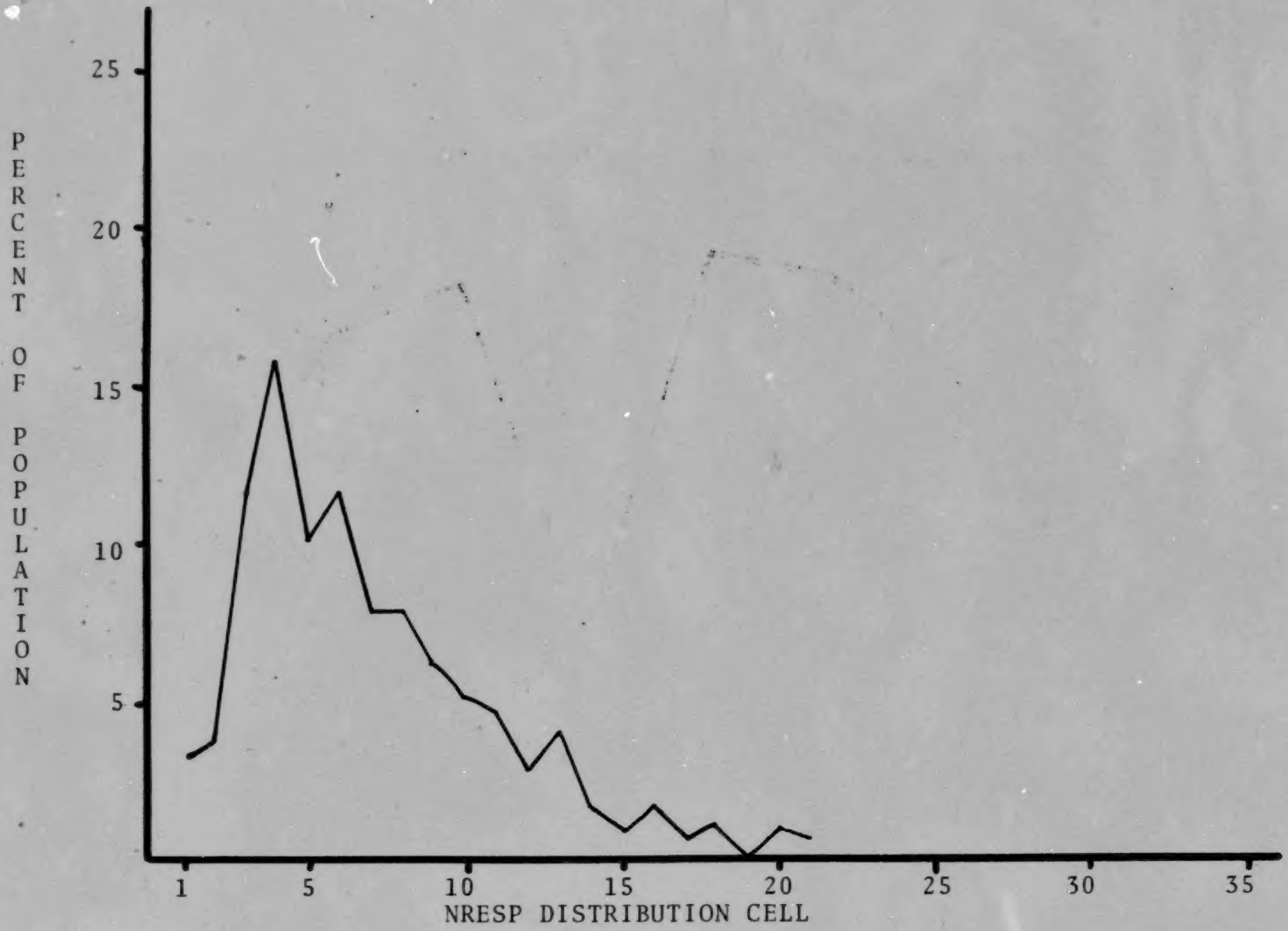


FIGURE 8. Distribution of numbers of responses (OF46).

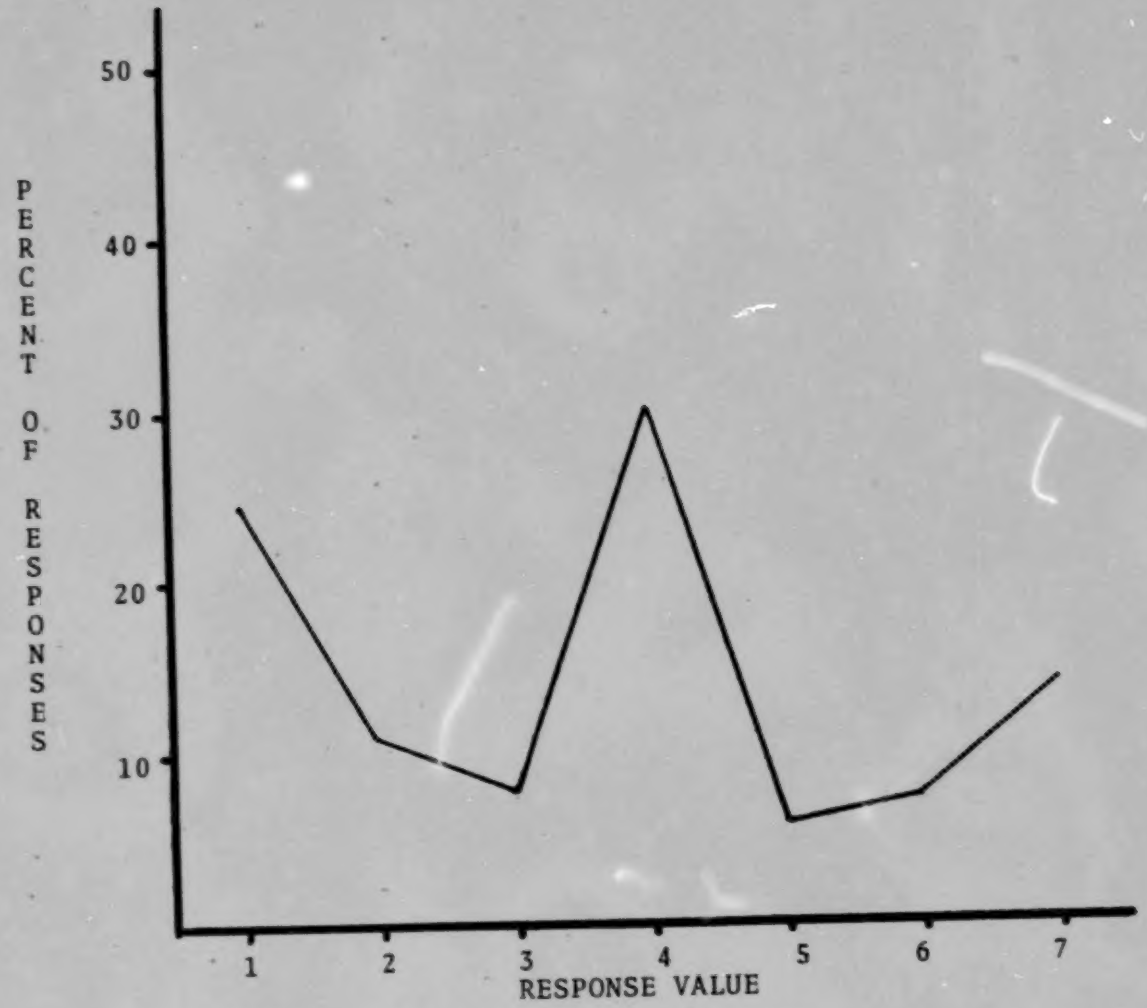


FIGURE 9. Distribution of response values (257X).

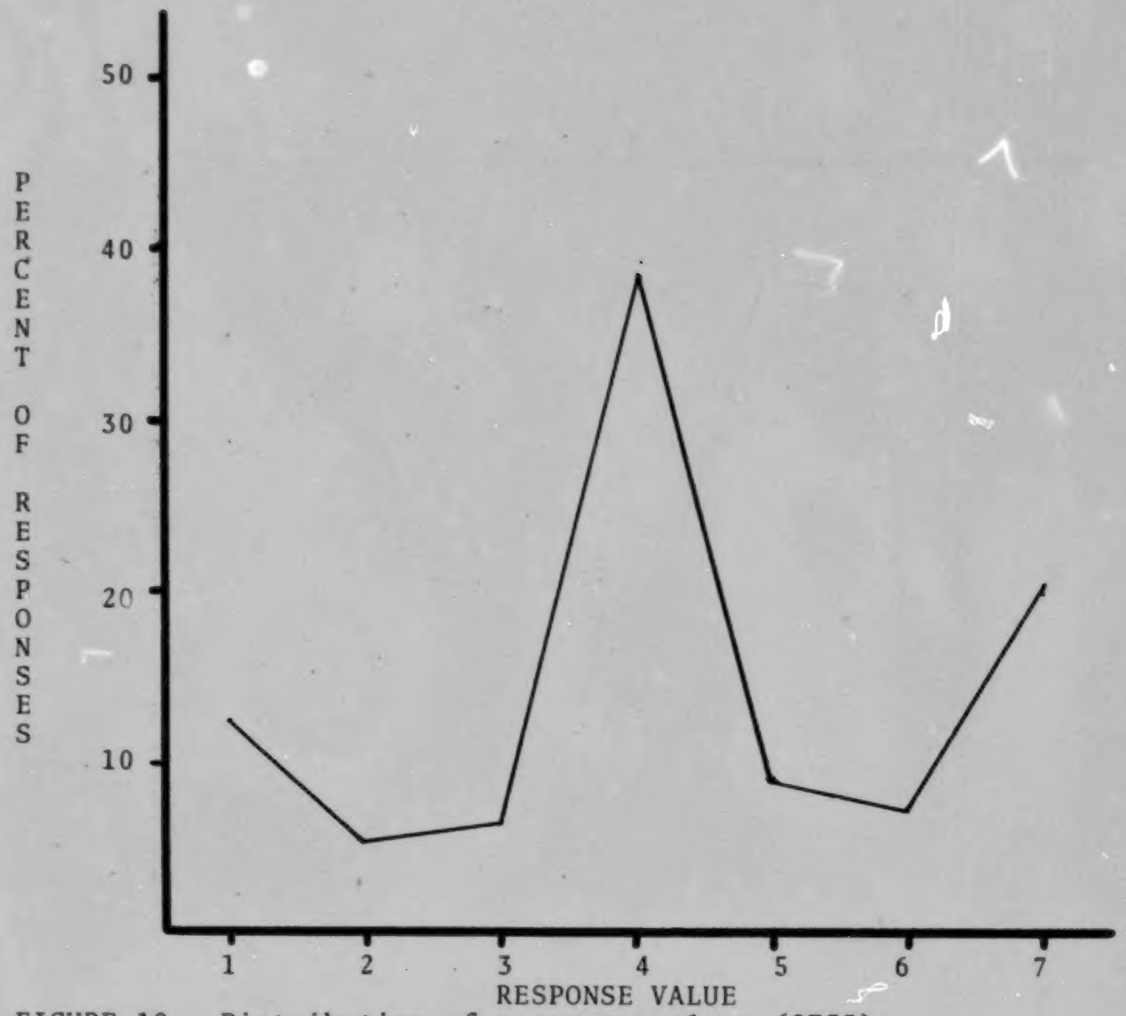


FIGURE 10. Distribution of response values (OF33).

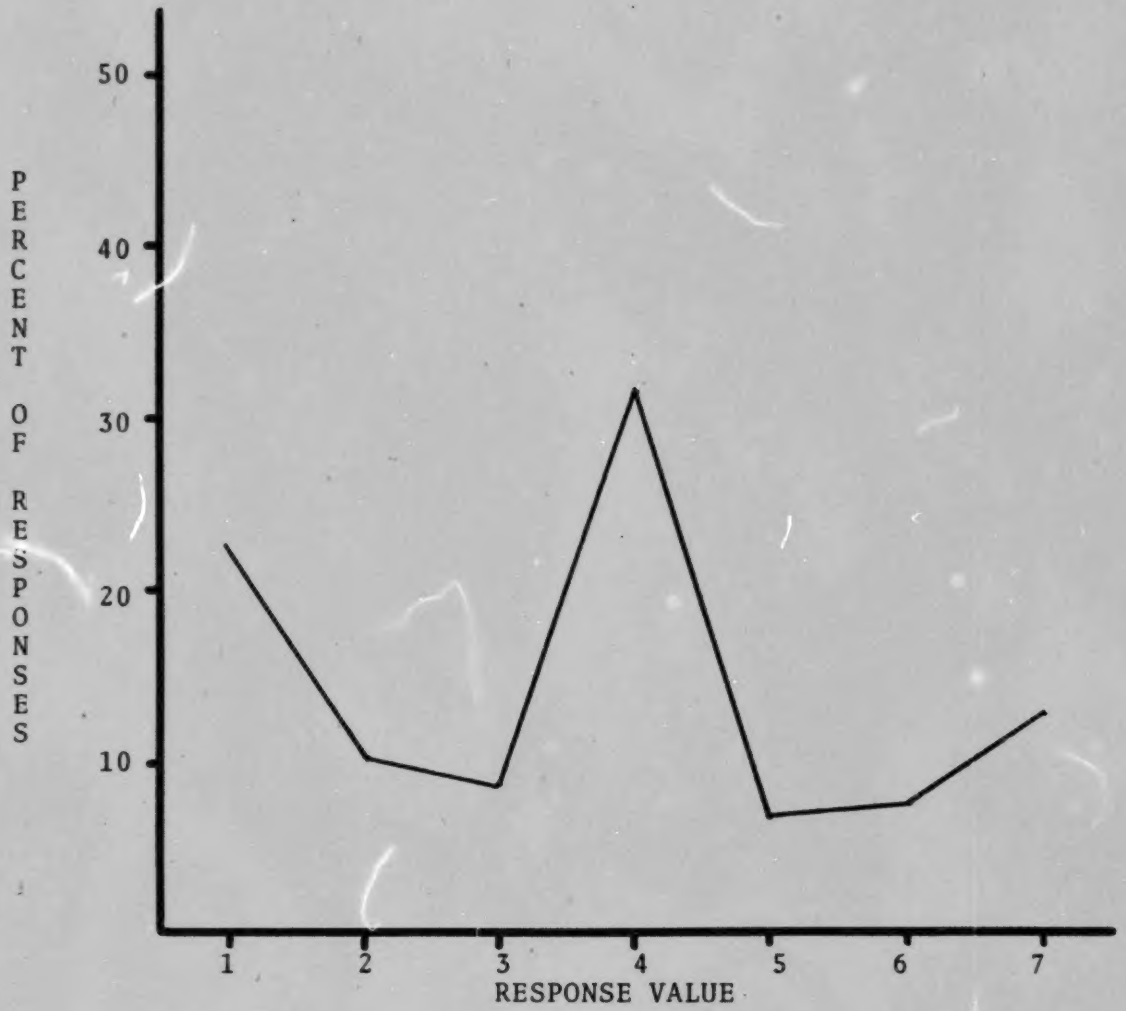


FIGURE 11. Distribution of response values (OF40).

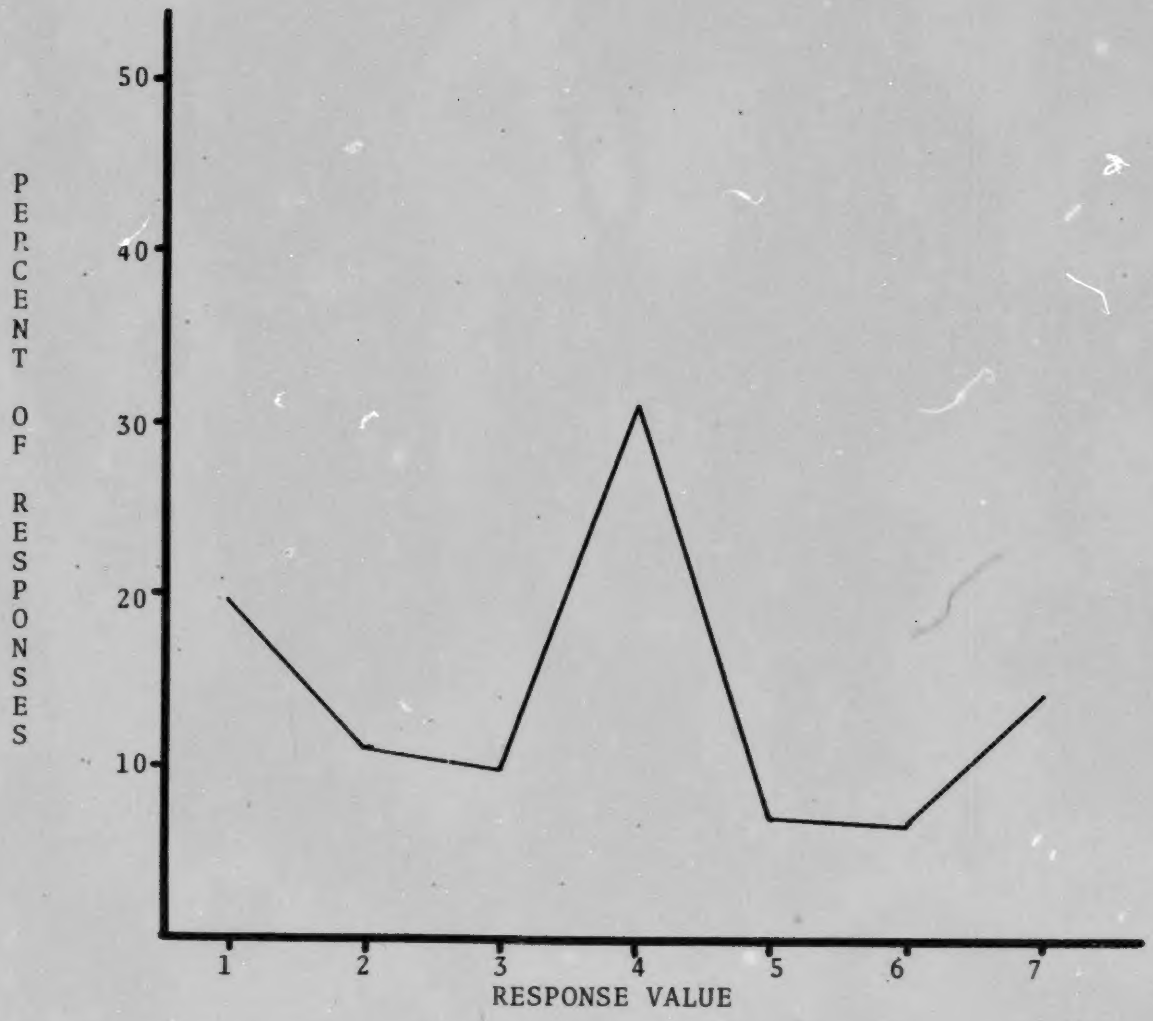


FIGURE 12. Distribution of response values (OF46).

"THE CONFLICTING DEMANDS OF GO-NO GO PERFORMANCE CRITERIA AND THE REQUIREMENTS FOR RANK ORDERING STUDENTS"

A paper delivered before the annual convention of

THE MILITARY TESTING ASSOCIATION
DR. WORTH SCANLAND
24 October 1974
Oklahoma City, Oklahoma

It is the purpose of this discussion to suggest a practical yet statistically sound solution to the problem which is posed by the differing requirements of some manpower managers and those who design, develop and deliver instructional programs using today's systems approach.

Although I speak solely from a Navy manpower manager perspective, I expect that the problem is or will soon be, universal. Managers who are responsible for the distribution of personnel, as well as those who are responsible for their advancement in rank, are usually interested in what we commonly refer to as class standing...that being a reflection of success in a school experience and therefore presumably a prediction of future success on the job.

On the other hand, those who design, develop and deliver instruction (which we can hereafter refer to as ISD) are not any longer interested in such categorizations.

The new concepts which modern ISD have brought to bear are several, but the two which are relevant to this discussion are:

(1) The use of task analysis as the data base for curriculum content, and

(2) The concept of mastery.

In the good old days we responded to a requirement to train radar technicians by giving an assignment for curriculum development to a petty officer or officer who was doing a tour of duty as an instructor, and he reached back into his personal store of experiences for the decisions as to what subject matter to include in the course, as well as for decisions as to how to measure how well his students had learned what he wanted them to know. Fortunately for training and for the students we do not perform ISD that way these days...we now send a team of job/task analysts to job sites to determine what it is that radar technicians must need to know and what they really do, how they do it, how they are supervised, where they learned each skill, how critical each skill is to job performance, and so on.

The outcome of this survey will be a set of skills which the new radar technician must have in order to perform his duties, and will include a list of those skills which he can best acquire in the schoolhouse setting. This set of skills does not include any "nice to know" items, and it does not leave any skill unattended to...in other words, the skill set is an expression of what must be learned, no more and no less. When this set of required skills is established, each is translated into specific behavioral objective terms - - that is, statements of required student performances which must be achieved or mastered. The acquisition of this minimum set of student objectives is the very essence of good ISD, because anything else is simply not cost/effective.

It is obvious, then, if we can accept the thesis expressed above, that each student in a modernly designed, developed and delivered instructional program is going to be required to achieve the same set of objectives as are all the other students, and that there are not any objectives left over for the smarter students to achieve which will earn them differentiation from their mates. To further confuse the issue of grading students, we are now delivering instruction in a continuum which does not allow for the formation of classes...students are continuously entering and leaving the program on an individual basis, and it is therefore not possible to rank order students in the customary "class" situation as we have in the past.

To summarize the underlying causes of our dilemma, we can say, then, that we must limit what skills are required of the students to those necessary to the good performance of their job or duty, and because they are necessary, we must require each and every student to demonstrate mastery of each and every skill...and on top of all that we are no longer dealing in finite classes or groups of students.

Is it reasonable to take the position that all students can achieve all of the objectives set for them...that is to say, can all of them achieve mastery of all the requisite skills? Happily the literature is replete with support for that contention, and I will refer you to the work of Carroll, Morrison, Bruner, Suppes, Skinner and Glaser, to name a few.

The concensus of these studies may be stated in this way...

If students are normally distributed with respect to aptitude for some subject matter, and all are provided with the same instruction in terms of quantity, quality and time, the end result will be a normal distribution on an appropriate measure of achievement, and the correlation between aptitude and achievement will be very high.

If, on the other hand, the students are normally distributed with respect to aptitude, but the instruction is tailored so as to be appropriate to each student, especially as regards time, 95 percent or more of the students may be expected to achieve mastery (Bloom, 1974). In fact Carroll (1963) states unequivocally that "aptitude is the amount of time required by a learner to attain mastery of a learning task".

It is not my purpose here, of course, to defend mastery as a form of objective achievement...that has already been well enough established. The point I do wish to make is that aptitude, mastery and time to learn are a set of interrelated factors whose interrelatedness provides us with a means of meeting the requirements of some manpower managers for rank ordering students. If, as Carroll has said, time to learn is the criterion of aptitude for mastery, we are in business. It only remains to design a logical, useful and practical method for applying this criterion, and that we shall attempt in the next few paragraphs.

One of the essential characteristics of good individualized instruction is that each student may proceed at his own pace through the instruction, having entered at a point which is commensurate with his current skills. Therefore we have a continuous stream of students entering, progressing through, and leaving the system, each at a different time and rate of progress. Thus there are no classes as we have traditionally known them, and we are therefore forced to consider the stream of students as a population from which we must establish individual differences. The gaussian or normal curve is a very handy tool for doing just this.

Most of us who are involved with measuring human traits, characteristics and performances are accustomed to the use of inferential statistics and the several tools with which it serves us. On the assumption that there may be, however, some of us who are not so familiar, I will ask your patience while I make a few basic statements about the gaussian curve and how we may use it to serve our current purpose.

Very simply stated, the gaussian or normal curve is a graphic representation of the likelihood that a certain event (such as completing a learning task) will occur when such occurrence is governed by the laws of chance. Fortunately for social scientists, human characteristics, such as aptitude for learning certain kinds of tasks, are the result of the random combination of genes, and therefore appear to be subject to the laws of chance, so we may, if this assumption can be accepted, apply the very useful mathematical properties of the normal curve to such events.

An essential consideration of the application of the normal curve to the solution of problems is that the events being studied are normally distributed amongst the population.

Let me confess right now that the populations of our military schools are not random samples of the universe, and that I am therefore suggesting a rather stretched assumption when I say that the students with whom we deal are indeed possessed with normally distributed characteristics. However, I will make the statement that they are sufficiently normally distributed to permit us, as social scientists attempting to find a reasonable solution to a problem, to use the normal curve as a tool.

As you know, the two most important characteristics of normal curve are the mean and the standard deviation, the former being the total number of events observed divided by the number of observations, while the latter is the variability of the events between observations. The baseline dimension of the graph represents numbers of events observed while the vertical dimensions represents the number of observations in which a certain event occurred. For example, the baseline might represent different times for completing an instructional program, with the minimum time at the right and the maximum time at the left. In this example, the vertical dimension could represent the number of students who completed the instruction in each of the time intervals measured.

By using a student sample of sufficient size to make the representative normal curve which we construct from the data one upon which we may generalize, we have obtained a method by which we may rank order students among a total number of students who have completed the instruction.

Because every student in whom we are interested has mastered the objectives set for him, the only remaining variable which we can measure is time, and this is of course very easy to do.

The procedure is straight-forward. Time to complete the required objectives is measured and noted for a randomly selected number of students... fifty will do, but two hundred would produce a better sample of the population.

With these data one can readily compute the mean time to complete for the sample, which is assumed to be the mean time for the population. Deviations from the mean for each of the sample students is computed, and then the standard deviation. With these parameters and a table of standard or "Z" scores, one can very readily answer any of the following questions:

(1) Did Seaman Jones complete the instruction in a time which is within the best ten (or any other) percent of the students who have taken this instruction?

(2) We wish to grant an "A" to all students who complete the instruction within the top fifteen percent of all the students' completion times...what is the "cut-off" time to make an "A"?

(3) What is the "cut-off" time for completion if we wish to grant a grade of "B" to every student who completes the instruction in less than the average time?

(4) Starting with those who complete the instructional program tomorrow, we wish to promote to the next higher rank the three hundred best qualified out of the next seven hundred who complete the course... what "cut-off" completion time will accomplish this?

(5) It has been determined that it is uneconomical to retain students in this instructional program for longer than N-days...what attrition rate will this attach to the program?

or we might turn that question around for a different purpose, saying,

(6) We can live with an attrition rate of seven percent...how long should we permit the slowest students to remain in the program?

There is an added dividend to this concept which is worth mentioning, I believe. If military students were paying for their training by the day or week, they would have a powerful incentive to progress as rapidly as possible. That not being the case, some other motivational incentives must be supplied...it is unrealistic to presume that the average student is going to give the learning task any more than the minimal effort required of him. Tangible evidence that rapid progress results in reward is going to pay off, and the system just suggested will help provide a means for doing that.

In summary permit me to be the first to identify the two key assumptions which underly the proposal set forth:

First, there is the assumption that time to learn is highly correlated with aptitude, and therefore should be an indicator of future success. I am very confident of that assumption and believe we can accept it for our purposes.

Second, there is the assumption that the population which is and will be attending to our pgoram of instruction is normally distributed with respect to aptitude for the training we are delivering. That is a shakey assumption, but not so shakey that I do not have confidence we can live with it, and there are statistical tricks by which one can normalize data if that is helpful to solving the problem.

If I have started you thinking about solving the problem of rank ordering students in a very modern instructional environment, my purpose has been well served.

References

Bloom, B. S. TIME AND LEARNING. American Psychologist, 1974, 29, 682-688
Carroll, J. B. A MODEL OF SCHOOL LEARNING. Teachers College Record, 1963, 64, 723-733.

SECTION XIV

SYMPOSIUM:

THE CARE AND FEEDING OF TEST CONSTRUCTORS

THE PERSPECTIVE OF A TESTING PSYCHOLOGIST WHO CONDUCTS ORIENTATION PROGRAMS FOR NEW TEST CONSTRUCTORS

WALTER BIRDSALL

The comments which follow applied to the Naval Examining Center when it was a separate entity. Some of the concepts and procedures still apply, others are undergoing modification, and still others fall under the category of "that was yesterday". Naval Education and Training Program Development Center is a young organization and an exciting one. For those of us who had long standing loyalties to our previous identities, we have amazed ourselves at the quickness we have blended and evolved a new allegiance.

A training program for examination writers just did not happen one day. Between the extremes of a four-month program to a three-day general orientation, many programs were tried -- many discarded. The Exam Center used E-7, E-8, and E-9 chiefs on a normal tour of shore duty as writers. We had about ninety such chiefs. Duty rotation dates were not fixed conveniently at a specified time. One month maybe no new cheifs reported in, the next month maybe six. Not all chiefs had the same length of shore duty. Annually, we would have between twenty-five to thirty new chiefs who would have to be trained. However, the advancement examination cycle is fixed. It begins one year prior to administration of the examinations. Each writer is responsible for developing an examination for each of the four paygrades within his rating. The development takes about five months. With two examination cycles a year, little time is left for training. When a new chief reported in, it was necessary to get him to produce as quickly as possible. A chief, well seasoned in examination writing may be transferred in the middle of a cycle and be replaced by an untrained man. Therefore, one factor dominated and controlled all other considerations. Timing. When to conduct a training program. The "when" would determine what to teach and to a large degree how it was to be taught.

Our solution was a departmental manual containing the procedural steps of examination development--a cookbook. Devoid of philosophical considerations and nicities, the manual states that first you do this, then you do that, etc. No attempt was made to include preferred style, techniques of writing rules of grammar. The intent of the manual was to familiarize the new writer with the steps necessary to accomplish his job. We found we had to expand one chapter beyond the intended scope--Statistics. This one chapter is a layman's approach to the broad body of statistics and also specifically a "how to" with those statistics which the writer must use in developing his product. Mrs. Elberta Soloman wrote this particular chapter which is a gem because it gets a difficult subject across without talking down to the reader. After a new chief reported in and was assigned to his division, he spent the first week reading and rereading this manual under the guidance of his division Education Specialist. From then on for the next two months, it became "hands on" training--implementing the manual, under very close supervision of the Education Specialist, to actually produce or finish producing an examination. Four times a year the chiefs who reported in during the previous quarter would meet every morning for a week with the department head and/or deputy to discuss any possible weaknesses in the manual and any problems in implementing the manual. As the time being aboard varied among the chiefs, the content of the meetings ranged from basic indoctrination to true critique. As there were no more than eight people at these meetings, the meetings were very informal and unstructured. An equally important purpose of these meetings was to reinforce the valuable contribution they were making. In the never ending production of examinations, the purpose of the examinations becomes elusive if not illusive. The examinations themselves become the purpose. The writer cannot lose sight of the true goal of properly manning the fleet with the best of the qualified. The task of examination development is a unique opportunity for the Exam Writer to make a significant and lasting contribution to the Navy of today and the Navy of the future--the man selected for advancement to Petty Officer Third Class today could well be a Master Chief Petty Officer in the Year 2000. This was the challenge presented in being assigned to the Naval Examining Center.

To maintain a check on the effectiveness of our training and to reinforce the concept of the uniformity of our testing product, we had a Qaulity Control Board. This board was a staff entity divorced from the Development Department. One of the major functions of this board was to assess the product of the department in terms of the standards of the department. This last point is important. It would be easy for such a board to evolve its own standards which could be different from those of the producers. Therefore, such a board must be vigilant that the standards used in assessing are truly the same as those of the producers. Any other posture is chaos, I assure you. I've been there.

The Quality Control Board must work closely with the Development Department in order that dual, or the appearance of dual leadership be avoided. In reviewing examinations, a board can ascertain the effectiveness of training of each division, can spot weaknessess that should be addressed. A good board functions as the major source of the effectiveness of the training of writers.

There was a weakness in our program--a weakness shared by anyone involved in developing examinations. Lack of resource material. In all of the books devoted to educational and psychological testing, there is little space devoted to actual item writing. The work of Robert Ebel is yet to be surpassed. However, his dicta are designed for people who are training for, or in, the field of Education and Psychology. Gronlund's work is a good extension of Ebel's, but still does not meet the need of a source for the non-professional.

We had started to develop two basic reference-text books. One book was to cover the basics of item writing; the other the grammar of questions. We had found that books on grammar do not quite meet the writing needs imposed by interrogatory rather than declarative style. However, the circumstances of consolidation have overtaken their completion.

As you can see, the program which we had was centrally controlled and centrally monitored.

Now we are in a new situation. Other products, along with and tantamount to the examination, are being produced. Whether to train for creating all products or to train for developing individual products is still to be determined. But we have an advantage over other new organizations in that we know that we must have a training program and we are aware of the problems that must be accomodated and/or solved.

The Perspective of a Military Supervisor of Test Constructors Who Are Temporarily Brought in from the Field to Write Tests

1LT DAVID VAUGHN

A typical test construction project at the USAF Occupational Measurement Center lasts about six weeks, and five tests are written for a particular Air Force specialty. These five tests include four Specialty Knowledge tests: two tests for promotion to E-5, one for active-duty personnel and another for Air National Guard/Air Force Reserve personnel and two tests for promotion to both E-6 and E-7, one for active-duty personnel and another for Guard/Reserve personnel. The fifth test is an Apprentice Knowledge Test. One use for this test is for assessing knowledge of active-duty personnel if they have experience in civilian jobs which are related to their military specialties. Approximately five subject-matter specialists (SMSs) are brought to the Center on temporary duty from throughout the Air Force for the duration of the project. These SMSs, who are senior noncommissioned officers (NCOs) and who are experienced in the specialty for which tests are to be written, provide technical knowledge in the subject matter of the specialty. A test psychologist, who is permanently assigned to the Center, works directly with the SMSs to make up a test construction team. These test psychologists, who are junior Air Force officers, like myself, or civilians, provide knowledge and experience in test construction policies and procedures. During the course of a test construction project, the expertise of a test psychologist and that of the SMSs are combined to produce the tests.

Every two test construction teams are assigned a review psychologist, also military or civilian, who provides a first-line quality control check. All test outlines, items, master test copies, and other material produced by a test construction team are checked by the review psychologist. A senior review psychologist, who is responsible for monitoring a large number of career fields provides a second quality control check. The key to good test development under this system is teamwork among the SMSs, the test psychologist, the review psychologist, and the senior review psychologist.

During the course of a test construction project, a number of tasks are performed. Current editions of the tests are reviewed, Air Force specialty descriptions are reviewed, and outlines for the new tests are developed. Old items are reviewed and revised when appropriate, new items are written, items are selected to appear on each test, and the completed tests are reviewed and finalized. With all these tasks to be completed during a six-week period, there is little time for formal training of SMSs in item writing. Some sort of training at the Center is necessary, however, since SMSs usually have little or no previous experience in test construction. How is this training accomplished?

Before SMSs arrive at the Center, they are mailed a packet of information concerning their temporary duty assignment. Included is a copy of the Handbook for Subject-Matter Specialists, which is published by the Center. The Handbook gives an overview of a test construction project. Included is information on developing the test outlines, old test items selecting and revising writing, new test items developing illustrated items, and reviewing and finalizing the test products. This Handbook serves as a useful reference for SMSs during a project, as well as helping them familiarize themselves with the test construction process before they arrive at the Center.

When SMSs first arrive at the Center, they are given a formal briefing. In this briefing, the Center missions are explained. The Weighted Airman Promotion System and the Weighted Airman Screening System are described and the relationship of the tests to these promotion systems is clarified. A brief overview of a test construction project is given in which the various phases of a project are outlined.

The Handbook and the formal Center briefing only serve as an introduction. The main responsibility for training of SMSs lies with the test psychologist. Although test psychologists have freedom in choosing the exact training procedures to be used, as a practical matter, most test psychologists use similar procedures. Most training is accomplished by short, informal briefings at appropriate points in the project, accompanied by extensive on-the-job training.

In the initial briefing given by a test psychologist, immediately after the SMSs arrive in the team room, SMSs learn what to expect during the project. The importance of the tests to the Weighted Airman Promotion System and the Weighted Airman Screening System is explained again. The responsibility of the SMSs for the final test products and the large impact that the tests will have on the future shape of the specialty is emphasized. SMSs are encouraged to always understand why they are doing something; since this will help them to do it better.

The first step of a test construction project is to review the current editions of the tests to be revised. The SMSs take the tests, and then go over each item with the test psychologist. The main purpose for this review is to delete obsolete or otherwise faulty items from scoring. However, this review is also an important training process. An SMS is able to get an approximate measure of his job knowledge and to gain insight into the actual experience of taking one of these tests from the standpoint of an examinee. An SMS is able to get a "feel" for what the tests are like, both in terms of format and content. At this point in the project, consideration can be given to what general sorts of changes should be made in the tests.

The next step involves development of a test outline and weighting of the outline areas to reflect the number of test items to be written on each outline topic. SMSs are briefed on the form of the outline and various guidelines in outline development, including a review of the specialty description. As the SMSs develop the outline, the test psychologist is present to offer suggestions and to help resolve conflicts. At this point, consideration is given to any recent or potential major changes in the structure of a specialty and to what reference material is available for use. The review psychologist and senior review psychologist may also discuss these matters with the team.

After the outline has been developed, items on the current version of the tests are reviewed for reuse. These items may either be reused exactly as they appear on the current tests, in which case they are considered verbatim items, or the items, may be reused with variations in the wording, in which case they are considered revised items. The training pattern of informal briefing and extensive on-the-job training is very clear during this part of the test construction process. SMSs are briefed on the various factors which should influence the decision to reuse and/or revise an item. Interpretation of the item statistics and test statistics is very important. For example, after I have explained the item statistics, I put some examples from the tests to be revised on my chalkboard and help the SMSs interpret the data to decide if the items should be reused, whether they should be revised, and how they should be revised. Processing procedures for old items are also explained. SMSs are then given all the old items to review. As soon as the decision to reuse an old item has been made and a revision, if necessary, has been written, the test psychologist looks at the item and provides immediate feedback to the team concerning any errors or good characteristics in their work. I think that the main training occurs during these discussions of items between the test psychologist and the SMSs.

The next step is the writing of new items. Some test psychologists present briefings on principles of new item construction together with the briefing on old item selection and revision. Although this is a reasonable approach, since old item revision requires knowledge of item-writing principles, I usually wait until the old items have been processed before I talk formally about item writing. In this new item-writing briefing, various rules and guidelines of item-writing are presented along with examples. The Center has published the SKT Workbook for Subject-Matter Specialists as an aid in teaching the principles. This publication contains items which illustrate various problems, along with a discussion of the problems in each item and a better version of each item, where appropriate. To illustrate cluing of the correct answer in test items, the Workbook contains an examination in Franzipanics. Franzipanics is a subject that nobody knows anything about, but all the items can be answered correctly from hints in the wording and format of the items. For example, one item reads:

> What is the primary purpose of the cluss in frumpaling?
> (a) To remove cluss-prangs
> (b) To patch tremalls
> (c) To loosen cloughs
> (d) To repair plumots

That answer (a), To remove cluss-prangs, is correct can be seen because it repeats cluss, a key word in the stem. Also, the correct answer is longer than the other choices and is the only choice that contains a hyphenated phrase.

Several test psychologists have prepared visual aids to help remind SMSs of the various item-writing guidelines while they are writing. Again, there is great emphasis on providing rapid feedback to the SMSs concerning their items. As items are written, the test psychologist reviews them and notes various problems and possible suggested changes. He then discusses certain items with the entire group of SMSs so that they may see, through practical example, what common mistakes to avoid and how to avoid them. Also, good characteristics of items are pointed out so that they may serve as models for other items. Various alternative item formats, particularly formats for illustrated items are discussed.

The same procedures for developing an outline, selecting old items, and writing new items are followed for each of the five tests. Although briefings are not usually given after the first set, the on-the-job training process of reviewing each item and discussing it with the item writer is followed throughout the entire project.

How well does this informal training procedure work? Can NCOs who are inexperienced in item writing be trained to produce a good test and, in addition, to produce five tests in six weeks? Yes, I think so. Of course, there is wide variation in how well a team learns to write items. I have had teams where after the first test, the changes that I made in items involved little more than corrections in spelling and grammar. On the other hand, I have had teams where I practically rewrote every item and, of course, had my revisions examined and approved by the SMSs. In all cases, I believe that our tests were of high quality and that they benefited from having been constructed by men who are fresh from the field and who know from first hand experience what their Air Force specialty is like today.

THE PERSPECTIVE OF A CIVILIAN TEST PSYCHOLOGIST WORKING WITH TEST CONSTRUCTORS

Carl R. Smith

US Army Enlisted Evaluation Center

Since the beginning of the Army's Enlisted Evaluation System in March 1958, the uses of evaluation results have constantly increased. In the beginning, the results were used solely to identify those soldiers eligible for proficiency pay. Since that time, promotion qualification, retention and reenlistment eligibility, special training eligibility, and other determinations have been based, in part, on evaluation scores. Today, the enlisted soldier's career largely depends on his performance as measured by various evaluation instruments. The written military occupational specialty (MOS) evaluation test is one of the main instruments used to evaluate soldiers.

Civilian psychologists in the Test Branch of the Enlisted Evaluation Center serve as principal contacts for MOS Evaluation Test Project Directors and their staffs at Item Writing Agencies (IWA) concerning matters relating to our Center and the Enlisted Evaluation System. We provide daily guidance via the AUTOVON telephone system and by mailings of FOUO and/or lengthy written information and guidance.

The enlisted evaluation written test cycle begins 14 months before the test administration month. This cycle begins with a form letter (subject: MOS Test Plans) sent from our Center to IWAs requesting test plans be submitted in accordance with a listing of MOSCs and a schedule of suspense dates for test plans in Inclosure 1. We discuss information requested in Inclosure 2 with the IWA test project director which includes proposed or recommended changes to MOS structures, enlisted soldier strengths in MOSCs, job descriptions, equipment items, status of Army publications (DA and non-DA) under revision, and classification of tests when other than FOUO.

These test plans are a very important step in the test development cycle. The Headquarters, US Army Training and Doctrine Command (TRADOC) sets and directs policy for Army service schools. They have directed schools to "systems engineer" all MOSCs for which those schools hold proponency. Relating to test planning, a systems engineered MOSC should contain a list of all duty positions and tasks.

Further, all such positions and tasks should be referenced to DA directives. Item writers, under the direction of the test project director at each IWA, should use the systems engineered study as the "keystone" of their work. Other regulations or documents to review for test planning should include, but not be limited to, job descriptions in AR 611-201/611-202 and the Military Occupational Data Bank (MODB) which lists task statements of an MOSC.

Ten months before a test administration month, a second request letter with two inclosures is sent from USAEEC to the IWA. Inclosure 1 is the schedule for submitting study references and item banks to USAEEC for a specific evaluation month or period. Inclosure 2 gives instructions for use of scheduled MOS test actions. Other inclosures attached to this letter are the MOSC test plans for the IWA for this evaluation period.

We develop a study guide from the references provided by the IWA. References are furnished to us for each skill level in the MOS. We merge a test plan, less numbers of items for areas, with each skill level (MOSC) reference list to make up this study guide. Also, certain standardized information about the purpose, scope, study material, study habits, training courses, score reporting, and correspondence study courses is included in this study guide.

Five fundamentals of item writer training which should be accomplished during FONECONs or TDYs with IWA test project directors and item writers should include:

1. Establish a good working relationship by being friendly, sincere, and enthusiastic about developing good quality test materials. Encourage item writers to do their best in learning their jobs and in developing their item writing abilities as rapidly as possible.

2. Set standards of performance which are fair and logical and will challenge their skills.

3. Frequently encourage item writers to demonstrate their skills within their capabilities.

4. Check as often as necessary to show that we (TPD and psychologist) are interested in the item writer's progress.

5. During TDY, give the item writers on-the-job training (special coaching) which may be required to meet the understood standards of performance.

The USAEEC psychologists take two TDY trips annually to each IWA. During these trips, they talk with the item writers, especially new item writers, to find out what is bothering them about the job, and then check these findings with their supervisor and/or test project director in the MOS test branch. Those training needs that fall within our responsibility should be noted and a brief training program developed to satisfy those needs. The psychologists conduct as much of this training as possible during the TDY week. If unable to conduct all the needed training during the working time available, assign the new item writers to the senior item writers who will serve as instructors.

Giving Instructions

Inform item writer instructors about the necessity to carry out accurate instructions to item writers as fast as they can understand them. If in doubt whether these instructions are understood, question the item writer. The minutes lost in making certain that instructions are understood may be regained many times, for then the item writers' mistakes will not have to be corrected at a later time.

Quality Guidelines

1. Set high quality standards and make certain that item writers understand them.

2. Get to know the writers. Advise them on how they can improve and give them recognition when they do.

3. Don't take less than the best. If an IWA sends in test materials that are not up to par, don't let it pass. Find out why. Maybe the item writer needs training or motivation is lacking. Friendly counsel may change the item writer's attitude. When the cause is found, some corrective action can be applied to it.

4. Keep communications moving. Give examples of the consequences of incorrect or incomplete communications, resulting in poor quality work. Help the item writers to recognize good quality by using specific examples.

5. Thoroughly inspect all materials submitted by the IWA to insure good quality and at the same time check the writer's performance.

6. Continually stress the necessity of an effective quality control program. When an item writer's idea is used, be certain to give the credit due him or her. Also, letters of appreciation should be written and presented to outstanding item writers by a responsible senior officer at IWA.

7. A good quality control program links every senior officer, every supervisor, and every item writer at the IWA. The links of such

a chain are strengthened by cooperation and coordination of efforts toward good quality test materials.

8. Maintaining good quality requires consistency in supervision. Psychologists, and the TPD, should assign item writers workloads that keep them busy. If the work assignment is too easy, the item writer develops an attitude of indifference.

9. Let the item writers know that they are the subject-matter experts, that we depend upon them, that we want to keep them informed, and that we want information from them that will improve test materials.

10. Item writers want to think that and to be told that they are doing a worthwhile job. The quickest way to kill initiative is to make them think that their work doesn't matter. The alert psychologist stresses the value of every job of every item writer that he or she works with.

11. Item writers like to work with a psychologist and TPD who assigns a job, trains them on how to do it, and then lets them go ahead on their own.

Guidelines to Building a Good Working Relationship Through Communications with IWA

1. Don't start talking before you know what you are going to say. Know what you want and plan the sequence of steps necessary to attain it. Study the assignment until it is so clear in your mind that you have no difficulty explaining it.

2. Identify your communications objective. Before you begin to communicate, ask yourself what you want to accomplish. Are you seeking facts? If so, decide what information you require so that you can ask intelligent, precise questions. Do you want to make a special assignment? Make certain you have analyzed the job thoroughly so you can explain it properly.

3. Don't do all the talking. Give other people a chance to tell what is on their mind. The only way you can convince people of your interest in them and your respect for their views is to hear their opinions and ideas. The psychologists who pay attention to what IWA people are saying learn more about them and more about their jobs. Remember that the purpose of communications is to build a foundation for "doing."

THE PERSPECTIVE OF A NON-COMMISSIONED OFFICER WHO IS ACTUALLY CONSTRUCTING TESTS:

ADCM J.R. HARRIS
INSTRUCTOR
U.S. COAST GUARD INSTITUTE
OKLAHOMA CITY, OKLA.

Before beginning my discussion, I would like to briefly describe my background. I have been in the Coast Guard for nearly 20 years. My specialty is aircraft maintenance, and for the last several years before coming to the Institute, I have been an aircraft maintenance supervisor. My duty tour at the Institute as a permanent duty instructor will be from 3 to 4 years.

My objective is to describe the problems that I have had as a new item writer. I will briefly describe our internal organization so that you can see where I fit into it and hopefully get a better "feel" for the problems I will mention.

Since the Institute is responsible for correspondence course development and servicewide exams, my time is divided between the two. I am required to prepare the E4 through E6 exams during July and August of each year and the E4 through E9 exams during February, March and April of each year. The rest of the year I develop correspondence training courses for my rating.

I begin an exam series by developing my individual exam outlines using as a guide the Coast Guard Enlisted Ratings Qualifications Manual that lists the skills required of each rating and pay grade. A warrant officer reviews the exam outlines for technical content and correctness. With the assistance of the exam evaluation staff and the test psychologist, I review the exam outline and statistical data on the previous exam to see where adjustments should be made in the revised exam. Final branch approvals are made by a middle grade line officer assigned as branch chief to each of the five branches - Deck, Engineering, Electronics, Clerical, and Aviation. When an exam is completed, it is reviewed by a GS-7 or 9 training specialist, who has a degree in English, journalism, or education. The completed exam then follows the same reviewing process as did the outline.

The adjustment problems for a new item writer vary from instructor to instructor, but the ones I have had are typical, I believe. One of the most difficult problems was being able to write questions with clear stems and plausible distractors. I knew basically what I wanted to ask, but I sometimes had difficulty translating that idea into a good question with plausible distractors. It has been my experience that the instructors who have the least amount of trouble in this area are the ones who come to the Institute with at least a fair background in basic English grammar. Our unit has a crash course in English grammar for new instructors to help them overcome this problem area.

When a new instructor reports aboard, he is dazzled by new terms and computer readouts that seem unfathomable at first. Such terms as skewness, kurtosis, P-value, discrimination, and reliability are confusing, to say the least, to a man who was supervising aircraft maintenance only a few weeks earlier. In my own case I developed about two series of exams before I felt that I really understood what the statistical data meant and how I could get the most out of it. I develop a series every 6 months, so I am speaking of one year of experience and approximately 1000 items. The Institute gives a very brief introduction to statistics to new item writers and an instruction booklet for reference. Actually using the statistics seemed to be the best teacher for me.

Another problem for a new instructor is determining the weight or the number of items to ask on each subject area of an exam. Rates in the Coast Guard are not nearly as specialized as they are in the larger armed forces. The wide range of skills covered by each rating can cause some men who are in the same rating but have differing assignment patterns to become more proficient in one aspect of their rating than in another. I have to guard against overweighting the sections in which I have had the most experience or the parts of my rate that appeal to me the most. In some skill areas, it is more difficult to write good questions, and a natural tendency is to shy away from these areas.

Creating a test item for a specific pay level gave me some difficulty and also some surprises when I first began. I was amazed at how some items performed. Some items that I thought should be easy for a particular pay grade turned out to be difficult, and vice-versa. I found that generally in my rating the lower pay grades seemed to do better on theory questions and the upper pay grades seemed to do better on the practical questions. These things can be learned only through experience as an item writer, since each rating has different requirements.

Keeping current on changes in the field can be a real problem for permanent duty instructors. In some ratings, such as electronics and aviation, changes occur at a rapid rate. It is very easy for the instructor who has been out of the field for a while to test on obsolete information. About the only correction factor we have for this is an occasional field trip or training school. Beyond that it is up to the instructor to try to keep up with field changes.

The fact that most of the specialties have only one representative at the Institute can sometimes be a problem. Since there is no one else in his specialty assigned, an item writer is forced to go with his own opinion at times. An inherent problem with this is the possibility of tunnel vision or "not being able to see the forest for the trees."

Another closely related problem is that since there is generally only one man per rating, quite often a transferring instructor departs before his relief is on board. The new man has absolutely no transition time to learn the new job. Since development schedules for examinations must be met, the man must sometimes do the best job he can with no training at all. Quite often an instructor must complete his predecessor's courses or exams without the benefit of notes outlining his predecessor's intentions or ideas.

A problem that affected the quantity of items produced was a lack of typing ability. Before coming to the Institute, I had taken a basic course in typing and had some typing experience. Although my typing speed has increased, at first it definitely limited the number of items I produced.

A major problem area for me as a new item writer did not relate to the mechanics of item writing but to an area that is generally ignored in discussing the problems of a new item writer - the problems of psychological adjustment. These problems, I believe, can and do have an impact on an item writer's job performance.

About all that the typical new item writer knows of the Institute before reporting aboard is that the correspondence study courses and servicewide exams for the Coast Guard are prepared there. I was somewhat apprehensive about going to a job for which I had not been trained and for which I might not have the aptitude to perform. I had had no formal schooling for quite a few years, with the exception of an occasional service training school. Therefore, there was some question in my mind as to whether or not I would be able to perform in a training environment.

I had a lot of difficulty adjusting to a desk job. Up to this point in my career, I had never had a job that required me to be at a desk for 8 hours a day. I think that most of the nonclerical rates find this to be somewhat of a problem.

Another problem that may seem rather trivial but is a real problem for some instructors is one relating to a loss of status. Since an item writer is a senior enlisted man, he has generally had subordinates working for him and quite often has been the top enlisted man in his particular work area. Because of the nature and the mission of a testing organization, nearly the entire work force consists of officers and senior enlisted men. A crew leader must convert back to a worker status, so to speak, at a point in his career and at an age that is not particularly conducive to changes of this kind.

A natural question to ask at this point would be, "In view of these problems mentioned, do our exams perform at an acceptable level?" This is a critical point since a man's standing on promotion eligibility lists is determined in large part by how well he does on the exam. Statistics provided by our test psychologist verify that the exams have performed acceptably. There were 92 different exams administered in our March 74 series. Forty-six of these had statistical populations less than 20, and 46 had populations of more than 20. Of the 46 with populations less than 20, 78% performed acceptably in the 3 areas of reliability, standard deviation, and mean. In the 46 exams with populations greater than 20, 95% showed statistical data above the acceptable level. Since statistical data is unreliable in the low population exams and half the exams are in this category, a great deal of reliance is placed on the instructor in our system. With an overall 87% exam efficiency, I would say that our instructors are overcoming most of their transition problems.

In conclusion, I have mentioned eleven problems that I consider the most difficult for a new item writer to overcome. Five of these problems are: (1) writing clear stems and plausible distractors, (2) understanding statistical data, (3) giving the appropriate weight to each exam section, (4) writing items for a specific pay level, and (5) keeping current with field changes. Two other problems relate to having only one instructor per rating: (1) "tunnel vision" and (2) no transition time for a new instructor. A lack of typing ability can be a problem for some.

Three of the problems I mentioned are psychological in nature. One of these is the apprehension involved in making a radical change to a job you are not sure you can perform. Also, there are the problems of adjusting to a desk job and the loss of status for some instructors.

Some of these problems can have a tremendous effect on the product the new item writer produces. Most of these problems are only temporary for most new item writers. By acknowledging that transition problems do exist for the new item writer, management can develop procedures and training programs that will help new men and will help in maintaining a more effective item writing staff.

"The Care and Feeding of Test Constructors"

By Lt Col George D. Oetting
Extension Course Institute
Gunter Air Force Station, Alabama

1. Introduction

Written tests permeate the military personnel system. From the time an individual takes tests to qualify for enlistment in the armed forces until he completes a full career, he is continually tested for classification, promotion and other purposes. Many of these tests are constructed by senior enlisted military personnel who know their jobs well and are highly proficient in the technical knowledge necessary to perform their operational duties. However, very few are professional experts in test construction procedures. Consequently, assignment to such duties may create special problems for newly assigned test constructors and the testing units to which they will be assigned.

These problems are the focus of this research. It is an initial attempt to explore this area within a number of testing activities of the armed forces. It includes the results of a telephone survey of testing units, a review of printed materials used to help orient new test constructors, and the detailed findings of a survey questionnaire of enlisted personnel assigned test construction duties.

2. Telephone Survey of Testing Units: To get an overall "feel" for the nature of testing programs operating in the military, a survey was conducted in March 1974 by telephoning major testing units in each branch of the military. Since military tests are used for many purposes, it was decided to concentrate on the following units which are deeply involved in tests used for promotion of enlisted personnel:

Military Branch	Unit Contacted
Army	Enlisted Evaluation Center
Navy	Naval Education & Training Program Development Center
Air Force	USAF Occupational Measurement Center
Marine Corps	Marine Corps Institute
Coast Guard	Coast Guard Institute

The following general areas were discussed with each unit:

a. Grade Level of Test Constructors: Most enlisted test constructors are in grade E-7, and above although others are occasionally used if necessary.

b. Assignment Status: All but the Air Force have test constructor personnel assigned on a stabilized tour of duty, lasting 2-4 years. Air Force personnel are brought in from operational units for 4-6 weeks temporary duty to write tests.

c. Use of Tests: Although these tests are used for promotion, their use in the promotion process varies. In some cases, the test score is the main factor in promotion; in others, test results are used as one factor in a promotion board and also for classification purposes.

d. Test Constructor Orientation: All units have developed an orientation program for newly assigned test constructors lasting 1-3 weeks. Quite often this is combined with on-going in-service programs throughout the tour of duty. In some cases, special workshops are conducted.

e. Test Design: All primarily use multiple choice items with 100-150 questions in each test. There are some variations, particularly in performance rating situations.

f. Testing Unit Organization: The overall promotion testing function is centralized in all branches of the military. However, those actually preparing the tests may be located at other units, such as in the Army; or the test constructors may be pulled in for testing duties, as done by the Air Force. All branches though, do have one unit designated as the final producer and controller of promotion tests.

3. A Review of Test Orientation Materials

There are many good reference books available in any academically respectable library on the principles and techniques of test construction. As part of this research, it was considered appropriate to take a look at the kinds of materials and activities used in the orientation programs of newly-assigned military test constructors. Therefore, the units contacted were asked to provide copies of the booklets, work sheets etc., that they used. Although bibliographic references were made to civilian texts, it was found that almost all material used was locally prepared. A review was made of materials received, and the following is a summary of the types of materials and orientation activities found in most units: (It should be emphasized that the summary cannot be considered comprehensive since it was not possible for the units to send all materials.)

a. Reference Materials Used:

1) Introductory information on testing philosophy and role of test constructor in the process.

2) General explanation of the overall sequence of local procedures to be used in test development.

3) General rules of good test construction with examples of poor and improved items.

4) Discussion with examples of the different types and formats of test items.

5) Suggestions on developing good distractors with examples of poor and improved versions.

6) Rules on use of illustrations and diagrams in tests with poor and improved examples.

7) General explanation of statistical concepts and terms used in evaluation programs.

8) Explanation with examples on the use of local test analysis statistics.

9) Explanation and examples of local administrative codings for test item cards, reports etc.

10) Procedures to be used in the safeguarding of classified and controlled test materials.

11) Check lists to be used as a final reference in the total sequence of test development.

b. Practice Activities:

1) Practice exercises in identifying errors in items and rewriting an improved version.

2) Practice exercises in identifying errors in either stem or distractors and rewriting an improved version.

3) Practice exercises in identifying and correcting errors in grammar, punctuation etc.

c. Particularly Interesting Orientation Activities:

The materials and activities listed above appear to be generally similar to the kinds of things mentioned in civilian texts. (With adaptations to local military requirements) This is quite natural since the rules for good evaluation are the same regardless of the testing environment. There were some interesting activities however, a bit different from the usual, that are worth noting:

1) The "Murder Board" - Several units used a technique in which a group of test constructors get together to review all items proposed for a particular test. These are projected or visually displayed in some manner to allow all personnel to "murder" the item through a vigorous exchange of questions, comments and critiques. If the item survives this session (with any changes needed) then it is considered for inclusion in the new test.

2) Performance-Oriented Tests - Almost all units expressed concerns about the need for making test items as realistic as possible rather than a simple regurgitation of isolated bits of information. Several units have attempted to do this by developing several multiple-choice questions measuring the thought process, procedures etc. required to perform a particular challenging, realistic military task. Quite often something like a completed form or a case situation is presented beforehand and used as the basis for answering a series of questions.

3) Nonsense Tests - In teaching principles of test construction, some units presented test exercises with nonsense words in test items which nevertheless can be answered if a person understands what rules of test construction are being violated. Another unit requires test constructors to write items violating a principle, and then rewrite the item in a correct format.

4. Initial Thoughts on a Survey of Enlisted Test Constructors: Having attended five previous conferences of the Military Testing Association, this investigator has listened to many papers and discussions covering the whole spectrum of testing activities from predictive models to performance rating techniques. Much has been presented on the use, design and analysis of tests. Unfortunately, very little has been said about the people who actually write these tests. This may be due to the fact that most conference attendees appear to be in supervisory positions and are required to concentrate on the "big picture" testing problems. This gap in research is the focus of this survey. It is an attempt to get some personal feedback from the enlisted troops at the working level of test construction.

5. Survey Design:

a. The survey format was a combination of open-ended questions and a structured rating scale. It was printed on one page (both sides) and designed to be completed in 15-25 minutes. (See atch 1)

b. Survey questions were designed to secure the following information:

1) Prior Preparation: Prior personal contacts with enlisted personnel assigned test construction duties revealed that many had little prior formal training or experience in this area. This question attempted to see if this is typical of military personnel assigned to this work.

2) Test Constructor Orientation: Most military testing units have some sort of orientation program to help newly-assigned test constructors learn their duties. This forced-choice question required the identification of the activity considered most helpful to new personnel.

3) Toughest Current Task: As with any job, some tasks prove more difficult than others. This forced-choice question required the identification of the one considered most difficult.

4) Selection of Test Constructors: After personnel have had some test construction experience and worked with others in this process, they should have a pretty good idea of the prerequisites for a successful test constructor. This question requires an indication of the three most important selection critieria in priority #1, 2 and 3.

5) Difficulty of Test Construction Tasks: This question required subjects to take a look at the total spectrum of major test construction tasks, and indicate the degree of task difficulty on a nine degree scale. Additional space was provided if a subject felt any major task was omitted. This list was developed by reviewing civilian and military test references.

c. A preliminary draft copy of the survey form was sent to personnel in the major testing activities of all branches of the military for review and comments. Particular care was taken to insure that potential respondees were familiar with testing terminology used.

d. Since this was a survey of military personnel, authorization had to be secured from the various military branches to administer it. This was secured in May 1974 and the survey was assigned USAF SCN 74-136.

6. Description and Size of Population Surveyed:

a. The population surveyed consisted of experienced senior enlisted personnel who are responsible for writing tests to measure the technical knowledge of enlisted personnel in their own career fields. Each branch of the service has somewhat different organizational structures and purposes in its testing program; however, they all have enlisted personnel performing this function. The selection of the sample was determined after consulting supervisors in major testing activities of all branches of the military. The following units were sampled:

UNIT	NUMBER SURVEYED
Army	
Institute of Administration	12
Military Police School	11
Signal Center and School	9
Transportation School	9
Navy	
Naval Education and Training Development Center	98
Air Force	
USAF Occupational Measurement Center (Subjects were from various USAF units)	49
Marine Corps	
Marine Corps Institute	27
Coast Guard	
Coast Guard Institute	24
TOTAL	239

7. Administration of the Survey:

a. The survey was sent out to the units to be surveyed in June 1974. Specific subjects were not pre-identified; this selection was left to the units. A letter of instructions was sent with the surveys explaining the criteria and precedures to be followed. (See atch 2) A cover letter was attached to each survey, explaining the nature of the survey to the subjects. (See atch 3)

b. After the survey forms were completed, the units involved collected them and returned then to the Extension Course Institute for analysis.

8. Limitations of Research: As work progressed in this research, it became evident that any conclusions needed to be tempered with the realities of the military testing world. These limitations should be considered when reviewing the summary of results:

a. Sample Surveyed: Although a sample size of 239 could be considered statistically respectable, it may or may not be typical of the total population of military test constructors. Time and personnel restraints prevented a sampling of all activities using military test constructors. Those selected were thought to be typical of the major kinds of testing activities in each branch of the military.

b. Subjects in the Sample: The great variety in the organizational structures established to administer testing programs created an equally great variety in the ways subjects were involved in the test construction function. This ranged from full-time test construction duties to a one-time temporary duty of several weeks to write a test. Consequently, the responses were from people with varying degrees of involvement in this process.

c. Open-Ended Questions: The usual problems inherent in open-ended questions emerged when analysis was attempted. Some questions, (particularly #1 and 2) proved too broad and subject to varying interpretation; some responses were unusable because they related to local problems not in the context of this research. A total of the tabulations for questions #1-4 revealed that some subjects did not provide any responses to these questions. As a result any sophisticated analysis of differences between various groupings of subjects was not attempted.

9. Results of Survey: The results of this survey are presented below in tabular form for each question. These additional comments may help in interpreting the results:

a. Questions 1-4, being open-ended questions, resulted in many kinds of answers. These were grouped as much as possible into similar-type answers, and a frequency count made. Some subjects made no responses, or responses not pertinent to this research; consequently the frequency tabulations are not a 100% coverage of the subjects sampled.

b. In addition to the frequency tabulation for question #4, a weighted frequency column was added, since subjects were asked to list three tasks in order of priority. It is felt that those rated as priority #1 should be given a higher weight, to get a more refined indication of cruciality. Therefore a simple weighting system, explained at the bottom of the tabulation, was added.

c. The results of the task ratings in question #5 were presented in two ways to better depict the information. One indicates the percentage (of all responses to a task) which selected a particular rating block; the other indicates the mean rating (using a 1-9 scale) and the range of one standard deviation, above and below the mean. Almost all subjects completed all responses on this scale, so lack of responses was no problem with this question.

(TABLES START ON NEXT PAGE)

"1. Prior Preparation: How much background did you have in test construction prior to your current assignment? Please indicate this by using the following two blanks:"

"a. Formal training: (Colleges courses, Military schools etc.)"

TYPE OF PREPARATION	FREQUENCY
None	115
Military school assignment	81
College courses in this area	18
Bachelor Degree	3
College teaching experience	2
Associate Degree	1

"b. Experience: (prior instructor duties, test writing duties etc.)"

TYPE OF PREPARATION	FREQUENCY
Prior military instructor duty	102
None	84
Prepared locally - used tests	34
Wrote military courses	14
Prior participation in test revisions	6
Served as troop commander	2
Acted as test administrator	1

"2. Test Constructor Orientation: While you were going through the orientation program for new test constructors, there were various activities designed to develop your skills. What one did you find most useful?"

ACTIVITY	FREQUENCY MENTIONED
Using printed reference materials	19
Participating in lectures and discussions	17
Actually writing test items	14
Reviewing and practicing language skills	13
Interpreting statistical test data	13
Reviewing rules for writing test items	12
Preparing testing objectives and outline	11
Working with experienced test writers	9
Studying examples of good and bad items	8
Analyzing test items written	7
Constructing stem of test item	7
Writing test item at appropriate level	4
Taking old tests which are to be revised	2
Writing a variety of test item forms	1
Learning how to research materials	1
Writing task - oriented questions	1
Constructing plausible distractors	1
Learning how to use dictionary	1
Using common sense in writing items	1

"3. Toughest Current Task: As you now are constructing tests, what do you find to be the single most difficult task?"

TASK	FREQUENCY MENTIONED
Writing plausible, parallel distractors	62
Writing items to measure specified proficiency level in a career field	33
Researching reference materials for testable material	27
Writing a testing outline/plan	19
Insuring test items are current	16
Constructing new test items	8
Writing items to meet statistical requirements	7
Interpreting test analysis statistics	6
Preparing comprehensive tests	5
Using proper English and grammar skills	5
Writing the stem of an item	4
Writing an easy question	4
Devising items without use of schematics	4
Coding items for administrative purposes	3
Arranging items within a test	1
Using behavorial objectives to make a test	1
Using creativity in constructing items	1

"4. Selection of Test Constructors: If you were in charge of selecting future personnel to perform your work in test construction, what would be the three most important criteria you would use. Please list in order of priority - #1 being the most crucial."

CRITERIA	FREQUENCY MENTIONED	WEIGHTED FREQUENCY*
Broad experience and knowledge in career field	196	514
Positive attitude, desire for this type of work	104	210
Knowledge of English, grammar	35	73
Ability to write clearly and simply	43	72
Broad educational background	40	71
Prior instructor experience	34	67
Prior training in test item writing	28	53
Professional military bearing and conduct	16	30
High military performance reports	13	26
Creativity, imagination, flexibility	17	25
Senior in rank - E-9 if possible	13	22
12-18 years experience - not too old	11	19
Common sense - objectivity	9	14
Ability to get along with others	8	11
Availability and retainability for work	8	11
Administrative experience and ability	5	11
Good reading comprehension	4	10
Up to date in own career field	5	5
Pass a personal interview	2	5
Able to take criticism	4	4
Ability to type	4	4
Has no personal problems	4	4
Able to understand statistical analyses	1	2

*Weighted Frequency determined by multiplying priority by a weight factor

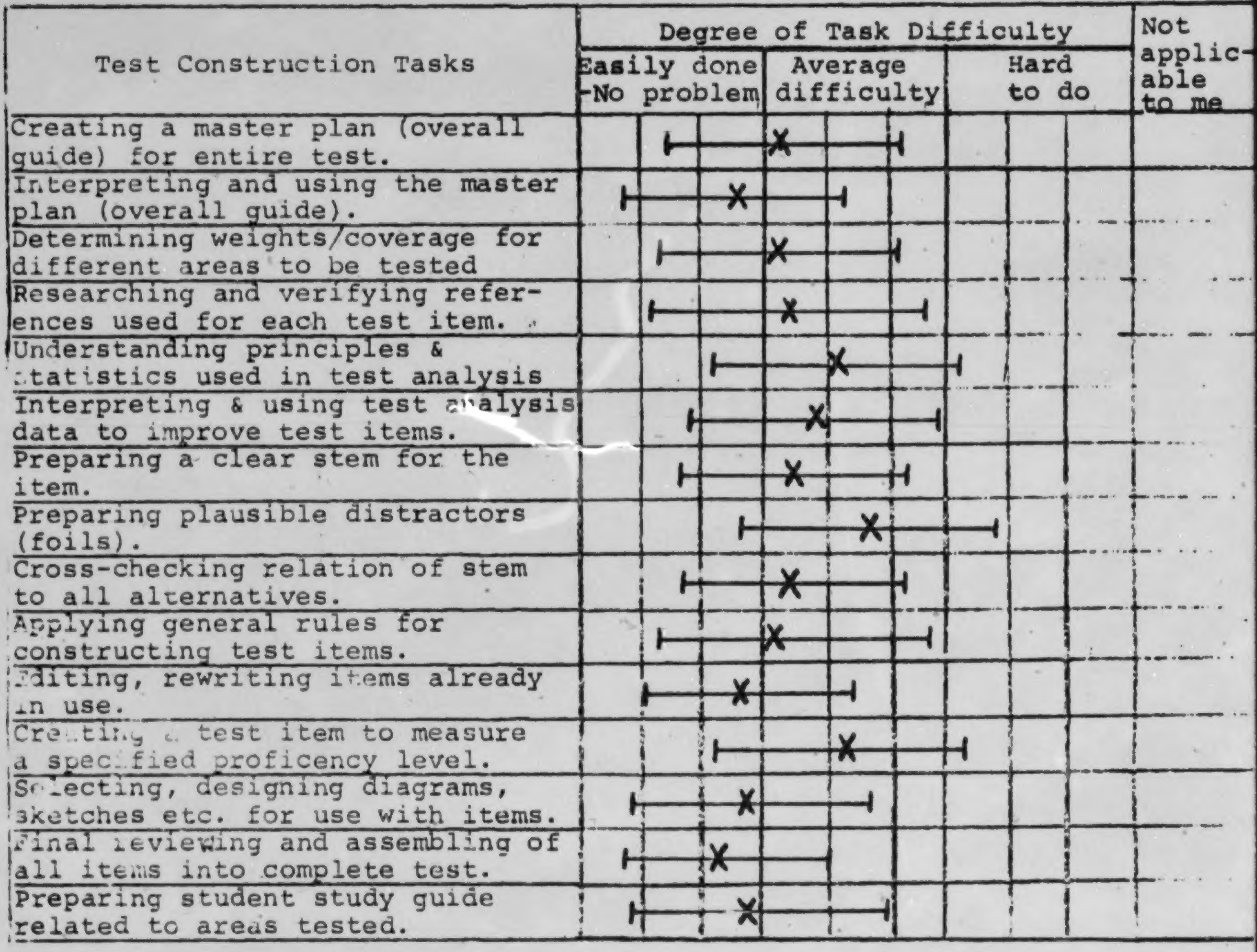
If priority is:		Weight Factor
#1		3
#2	Multiply by	2
#3		1

"5. Difficulty of Test Construction Tasks: In your current assignment, you soon discovered there are many different tasks involved in test construction. We have tried to identify the major ones in the list below. We would like your opinion on the degree of difficulty in performing each of these tasks, once you really got started in constructing tests. Please indcate this through a check for each task. (There are 3 blocks in each major division so you have a 9 degree scale available for each response)"

Test Construction Tasks	Degree of Task Difficulty									Not applicable to me
	Easily done -No problem			Average difficulty			Hard to do			
Creating a master plan (overall guide) for entire test.	7	11	10	24	20	11	9	1	2	5
Interpreting and using the master plan (overall guide).	13	14	23	18	16	6	6	1	1	2
Determining weights/coverage for different areas to be tested	8	10	14	22	15	13	10	3	0	5
Researching and verifying references used for each test item.	10	8	18	17	12	16	10	4	4	1
Understanding principles & statistics used in test analysis	6	4	10	20	16	12	17	6	5	4
Interpreting & using test analysis data to improve test items.	4	6	14	18	16	16	10	6	5	5
Preparing a clear stem for the item.	5	6	13	29	17	14	10	3	1	2
Preparing plausible distractors (foils).	4	3	7	15	10	16	23	9	10	3
Cross-checking relation of stem to all alternatives.	5	7	16	21	24	13	8	3	1	2
Applying general rules for constructing test items.	7	9	20	25	19	8	7	2	0	3
Editing, rewriting items already in use.	12	11	24	22	16	8	5	1	0	1
Creating a test item to measure a specified proficiency level.	5	6	6	18	19	11	17	7	6	5
Selecting, designing diagrams, sketches etc. for use with items.	13	12	19	18	12	6	7	2	1	10
Final reviewing and assembling of all items into complete test.	17	12	18	24	14	6	3	0	0	6
Preparing student study guide related to areas tested.	9	6	13	8	10	4	5	1	2	42

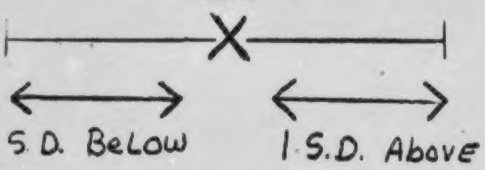
The numbers above indicate the percentage of total responses which selected a particular block on the rating scale. The highest percentage for each task has been circled.

"5. Difficulty of Test Construction Tasks: In your current assignment, you soon discovered there are many different tasks involved in test construction. We have tried to identify the major ones in the list below. We would like your opinion on the degree of difficulty in performing each of these tasks, once you really got started in constructing tests. Please indcate this through a check for each task. (There are 3 blocks in each major division so you have a 9 degree scale available for each response)"



Test Construction Tasks
Creating a master plan (overall guide) for entire test.
Interpreting and using the master plan (overall guide).
Determining weights/coverage for different areas to be tested
Researching and verifying references used for each test item.
Understanding principles & statistics used in test analysis
Interpreting & using test analysis data to improve test items.
Preparing a clear stem for the item.
Preparing plausible distractors (foils).
Cross-checking relation of stem to all alternatives.
Applying general rules for constructing test items.
Editing, rewriting items already in use.
Creating a test item to measure a specified proficency level.
Selecting, designing diagrams, sketches etc. for use with items.
Final reviewing and assembling of all items into complete test.
Preparing student study guide related to areas tested.

The figures above depict the mean rating and standard deviation for each task:



"(If you feel there are additional major tasks not mentioned above, please write these in below and indicate degree of difficulty.)"

ADDITIONAL MAJOR TASKS	Degree of Task Difficulty								
	Easily done -No problem			Average difficulty			Hard to do		
Preparing understandable training objectives									X
Relating test items to rating qualifications							X		
Meeting format differences between course writers						X			
Systems engineering									X
Measuring different enlisted grades within same skill level							X		
Formulating objectives				X					
Writing clear student instructions						X			
Overall management of testing program							X		
Typing material for review							X	X	XXX
Limited reference materials									XXX
Keeping updated on career information							XX	X	
Preparing item with limited references								X	
Coding and administrative procedures				X				X	XXXX
Ability to use reference publications				X					
Analysis of tests with small population									X

The additional major tasks listed above are those added by subjects, which were not preprinted on the task list. In terms of the total number of subjects, they are not considered to be significant. Most of those receiving more than one "X" came from the same unit and appear to be more of a local gripe about final administrative requirements.

10. Summary of Survey Results: Keeping the limitations in mind outlined above, the following is a brief, subjective summary of the findings:

a. The majority of subjects have had no formal training in the test construction area. Those that have, obtained it through prior experience as a military instructor. Some have had experience in writing locally used tests or through writing military courses.

b. Subjects had divergent opinions on what was the most useful activity in orientation programs. Printed materials prepared by the local unit proved most useful. Also rating high were participation in group discussions/lectures and actually writing test items. Many felt the review of needed language skills and the interpretation of statistical data were valuable.

c. By far, the toughest task was felt to be writing plausible, parallel distractors for multiple-choice items. Two other tasks rating high in difficulty were writing items to measure a specified proficiency level within a career field, and finding suitable references with testable material.

d. Subjects clearly indicated that broad experience and knowledge in a career field was the most important criteria in selecting future test constructors. A positive desire and attitude for this type of work was considered crucial. Communication skills and a good educational background were also rated fairly high.

e. In rating all tasks on a difficulty scale, most tasks were considered in the average difficulty range. There were gradations however, with preparing plausible distractors, measuring specific proficiency levels and understanding test analysis statistics being rated most difficult. The easier tasks were interpreting the testing plan, editing old items, and using diagrams and sketches with items. Most felt that the task list covered the important areas of their work.

11. Potential Uses of Survey Information: It is hoped that the results of this survey will prove useful to other personnel involved in test construction. These are some of the possibilities:

a. Review of Orientation Program: The results indicate there are differences in the perceived difficulty of the various tasks in test construction. A review of your program may be appropriate to see if there are sufficient orientation activities to insure proficiency in the more difficult tasks such as writing plausible distractors and writing items to measure different proficiency levels.

b. Development of Orientation Program: If you are starting or revising an orientation program, the task list may help to provide some structure in planning the necessary activities. Most respondees seemed to feel that this list covered the major tasks involved in their work.

c. Improvement of Orientation Critique: The kinds of questions asked may prove useful in any critique forms you use to evaluate the effectiveness of your own programs. These could be easily modified to meet local needs.

d. Selection of Military Test Constructors: The criteria recommended by the military test constructors themselves are probably a realistic, honest appraisal of the qualities needed for this type work. They should be considered as you review personnel records and interview potential assignees.

e. Basis for Further Research: This survey was planned as an initial exploration of this subject. As such, the research design was not very sophisticated in terms of statistical treatment. Conclusions drawn on the basis of percentages and response tabulation cannot be considered as "proof positive". However, it is a start, and might be useful in developing more definitive research questions in this area.

f. Stimulate Further Survey Research: It was abundantly evident while conducting this research, that a great deal of time, money and personnel are devoted to testing activities within the Department of Defense. Many of the same kinds of functions are performed by these units. The survey approach used in this research could be applied to many areas of mutual interest. The pooling of research, experience and techniques in research reports could provide a valuable reference for others involved in the same area. It could stimulate interest develop personal contacts and help prevent duplication of effort.

SURVEY OF TEST CONSTRUCTORS
(USAF SCN-74-136)

1. Prior Preparation: How much background did you have in test construction prior to your current assignment? Please indicate this by using the following two blanks:

a. Formal training: (College courses, Military schools etc.)

b. Experience: (prior instructor duties, test writing duties etc.)

2. Test Constructor Orientation: While you were going through the orientation program for new test constructors, there were various activities designed to develop your skills. What one did you find most useful?

3. Toughest Current Task: As you now are constructing tests, what do you find to be the single most difficult task?

4. Selection of Test Constructors: If you were in charge of selecting future personnel to perform your work in test construction, what would be the three most important criteria you would use. Please list in order of priority - #1 being the most crucial.

#1.

#2

#3

Attachment 1 (Front)

(Please Turn Over)

5. Difficulty of Test Construction Tasks: In your current assignment, you soon discovered there are many different tasks involved in test construction. We have tried to identify the major ones in the list below. We would like your opinion on the degree of difficulty in performing each of these tasks, once you really got started in constructing tests. Please indcate this through a check for each task. (There are 3 blocks in each major division so you have a 9 degree scale available for each response)

Test Construction Tasks	Degree of Task Difficulty									Not applicable to me
	Easily done -No problem			Average difficulty			Hard to do			
Creating a master plan (overall guide) for entire test.										
Interpreting and using the master plan (overall guide).										
Determining weights/coverage for different areas to be tested										
Researching and verifying references used for each test item.										
Understanding principles & statistics used in test analysis										
Interpreting & using test analysis data to improve test items.										
Preparing a clear stem for the item.										
Preparing plausible distractors (foils).										
Cross-checking relation of stem to all alternatives.										
Applying general rules for constructing test items.										
Editing, rewriting items already in use.										
Creating a test item to measure a specified proficiency level.										
Selecting, designing diagrams, sketches etc. for use with items.										
Final reviewing and assembling of all items into complete test.										
Preparing student study guide related to areas tested.										

(If you feel there are additional major tasks not mentioned above, please write these in below and indicate degree of difficulty.)

DEPARTMENT OF THE AIR FORCE
EXTENSION COURSE INSTITUTE (AU)
GUNTER AIR FORCE BASE, ALABAMA 36118

SUBJECT: Survey of Test Constructors

TO:

1. We greatly appreciate your willingness to administer this survey to people in your unit. The results of this survey combined with a review of service-wide procedures in this area will hopefully result in a useful study for everyone concerned. A report of this research will be presented at the Military Testing Association Conference in October; you will also be sent a copy of the study about the same time.

2. In administering the survey please use the following guidelines:

a. It should be given only to senior enlisted military personnel whose duties include writing test items on a regular basis.

b. Try to select people who still have a pretty good idea of the initial problems facing people learning how to write good test questions.

c. For your unit, we would like to survey   people if possible. A few extra copies are also included.

d. Please explain the purpose of the survey and assure the respondees that their answers will be kept confidential and combined with all others being surveyed. The results in no way will be related to any individual or to any testing unit.

3. The survey is a short one - it should only take 15-25 minutes to complete both sides of the one page. We would appreciate your conducting the survey as soon as possible and return the completed forms in the attached envelope by   This will give us time to tally and analyze the results and prepare the report for the conference.

4. This survey has been reviewed and approved for use in all branches of the Military. (Survey control Number: USAF SCN-74-136).

5. Thanks for your help. If you have any questions give me a call on Autovon, 921-4391/4392.

GEORGE D. OETTING, Lt Col, USAF
Chief, Educational Systems Branch

1 Atch
Surveys -   Copies.

Strength Through Knowledge

Atachment 2

DEPARTMENT OF THE AIR FORCE
EXTENSION COURSE INSTITUTE (AU)
GUNTER AIR FORCE BASE, ALABAMA 36118

DEPARTMENT OF DEFENSE

SURVEY OF TEST CONSTRUCTORS

Would you please take a few minutes to answer some questions about your experiences while learning how to write good test questions. As you undoubtedly already know, it is a difficult job to create an effective test. Maybe you felt somewhat like I did when sitting down to write my first test. I had taken hundreds of tests over many years of schooling and even had some courses in tests and measurement. But these experiences really didn't help much when I shifted from the role of test-taker to test preparer.

This problem is the focus of the attached questionnaire. We are interested in getting reactions of military personnel like yourself who have been assigned the tough but very important task of creating tests to measure the technical knowledge of military personnel. We are planning to query people from the Army, Navy, Air Force, Marines and Coast Guard who are involved in this work. Your reactions will be combined with all others and presented in a special symposium at the 1974 meeting of the Military Testing Association, hosted by the Coast Guard in Oklahoma City. Copies of the research will be sent to all units involved in the survey in October of this year. If you want a personal copy please drop me a note and I'll send you one.

This questionnaire has been reviewed and approved for use in your branch of the military. (USAF SCN-74-136)

Thank you for your assistance

George D Oetting

GEORGE D. OETTING, Lt Col, USAF
Chief, Educational Systems Branch

Attachment 3

SECTION XV

GENERAL SESSION

Research Report No. 74-6

October 1974

THE DEVELOPMENT OF AN INSTRUMENT FOR STUDENT EVALUATION OF COURSE AND INSTRUCTOR

Lieutenant-Colonel G.J. Carpenter

The research for this paper was supported in part by the Defence Research Board of Canada, Grant Number 9465-12. The views expressed are those of the author and do not necessarily reflect official views or policy of the Royal Military College or the Department of National Defence.

Department of Military Leadership and Management
Royal Military College of Canada, Kingston, Ontario
Department Head: Lieutenant-Colonel G.J. Carpenter

October, 1974

THE DEVELOPMENT OF AN INSTRUMENT FOR STUDENT EVALUATION OF COURSE AND INSTRUCTOR

LCol Gerald J. Carpenter
Royal Military College of Canada
Kingston, Ontario

ABSTRACT

There are many course evaluation questionnaires available, but a common shortcoming of such off-the-shelf articles is that they do not fit the customer. The reason is that the criteria for good versus bad instructors or courses must be chosen from a limited questionnaire. Analysis of such a questionnaire gives a rank ordering of characteristics which are pre-selected by the questionnaire author, and though such a list may be good for his training institution, it may not be what is required by others.

An approach to ensure relevant items on course evaluation instruments is to develop them through the critical incident method as described by Flanagan (1954). The result is, to be sure, a limited questionnaire but its content is tailor made, not off-the-shelf.

A project to develop a Professor and Course Evaluation Form was undertaken at the Royal Military College of Canada, using the critical incident technique. The method by which data were gathered for the study, items developed, and some results from subsequent use are reported.

Staff reaction to assessment and the use which can reasonably be made of course critique data are discussed.

THE DEVELOPMENT OF AN INSTRUMENT FOR STUDENT EVALUATION OF COURSE AND INSTRUCTOR[1]

LCol Gerald J. Carpenter

Introduction

In the military we are all familiar with the end of course critique. It may range in format from a narrative report of items discussed in a class bull-session, through lists of open-ended questions, to multiple-choice questions on specific aspects of the course.

The narrative report ends up being little more than one man's opinion, no matter how hard the writer tries to reflect the feelings of the group. The open-ended questions result in such a mass of conflicting data that the professor is left with a virtually impossible task in data reduction. The objective item question in either multiple choice or Likert format appears to be the answer - but it is only an appearance. It has its own set of problems. What it gains in focus it loses in breadth of field.

An objective test of 20 items covers a fairly narrow spectrum, and may miss those things that you really need to know (i.e., that the students really want to tell you). A test of 100 items that covers everything you can think of may still miss an important aspect of teaching, and anyway 100 items is too long for this kind of instrument. Who is this guy who makes up evaluation form items anyway, and how can he presume to cover the ground that students find important? We want a student assessment of professor and course, but someone else has chosen what the criteria for assessment, the items, are going to be.

That is the problem to which this paper is addressed, a problem of not only focus and field, but who gets to point the instrument.

---

[1] Cadets L.T. Hunt and T.L. Hunt collected the data, sorted incidents and named the critical requirements for the instrument described in this paper.

Literature Review

I cannot review the literature exhaustively, but I wish to bring a few studies to your attention. Holmes (1971) used about thirty items designed to assess teaching ability; items such as "instructor well-prepared", "revealed enthusiasm for teaching", and "returned assignments promptly" were included. Item intercorrelations were computed, and a principal axis factor analysis carried out. From the factors that emerged the four best were chosen, and items that loaded on the respective factors were assembled into scales. I am sure you can see how circular such a procedure is. The refinement of getting four namable factors doesn't change the fact that the items form a closed system.

Mueller et al (1971) used the Gadzella Quesionnaire (Gadzella, 1968) to obtain college student views of an ideal professor. The Gadzella Questionnaire contains 25 positive statements about professor behaviour, and the student selects the 5 most important and the 5 least important statements. Some examples of the statements are "has a thorough knowledge, both basic and current, of the subject he teaches", and "has a deep interest in and enthusiasm for the subject he teaches". Analysis consists of weighting the 10 chosen statements from +5 (most important) to -5 (least important). Finally, the items are ranked on the basis of the total weight received from the group.

Again the shortcoming of choice from a "limited questionnaire" is evident. The professor can only be described in terms of a check list whose appropriateness is not known. The authors concede this limitation with the statement "The meaning of the results, thus, is in the context of the characteristics presented (p 167)."

The Problem

I hope the preceding information helps to clarify the problem which I will now state as one of domain validity.

The domain has to encompass items which the students have stated are the ones important to them in describing instructor behaviour. At the same time the questionnaire has to be focused, and less than 30 items in length, or the users are likely to tire before it is completed.

Method

The method chosen to meet these seemingly incompatible demands was an application of Flanagan's Critical Incident Technique (Flanagan, 1954). This technique is widely used in job analysis, and was the method chosen for developing the "critical requirements" of leadership, a Canadian Forces study reported previously to this body by Otke (1969). The heart of the method is to obtain first hand reports of satisfactory and unsatisfactory execution of the task under review. In the case at hand the interviews were conducted among third and fourth year (i.e., juniors and seniors) cadets. It was felt that a first or second year man might not yet have sufficient experience in the university classroom to be able to relate valid incidents. The stimulus questions used were:

1. Think of the most interesting class you have had in your 3 or 4 years at Military College.
   Answer the following questions:
   a. What did the professor do to make the class interesting?
   b. What was there about the class itself (other than the professor) that made it interesting to you?
2. Think of the least interesting class you have had.
   Answer the following questions:
   a. What did the professor do to make the class uninteresting?
   b. What was there about the class itself that made it uninteresting to you?

Each interviewee gives only one response to each of the above questions. Data reduction consists of sorting incidents (answers) into piles based on answer content similarity. They must not be sorted into preconceived categories. After sorting, the answers in each pile are reread and a category name is assigned based on the content of the answers. The categories developed in this way are called "critical requirements".

Critical requirements developed from the second stimulus question (uninteresting) are examined to see if they are merely the obverse of a positive critical requirement. Many will be eliminated on this ground, but a few "bad" examples will come through for which the opposite did not receive mention as a "good" incident. For example, under 2.b., several students responded "poor, outdated reference material". No one, however, said that a class was enhanced by "good, up-to-date reference material" though it is known that the majority of professors use only that kind.

If the number of critical requirements is excessive (over 30), the possibility of combining similar piles should be examined. Reliability checks can be made by having independent sorters develop categories. Differences are resolved by mutual discussion which also serves as sorter training.

Finally, the critical requirement names are cast into a form suitable for use in an assessment questionnaire. The form chosen at RMC is a statement of the critical requirement and a five-point Likert scale. The twenty-two items that appear on the form consist of 15 statements about the professor and 7 statements concerning the course. Responses are made on a mark sense card and a programme to tabulate and summarize the data has been written.

A list of the 22 critical requirements from this study are displayed in Table 1.

Insert Table 1 about here

It is often assumed that frequency of reporting a particular critical incident is a measure of its importance. Perhaps it is better to call it salience rather than importance. The critical requirement ranks based on frequency of report are shown in the table.

Results of Use of the Form and Discussion

I would like to say that the quality of instruction at the Royal Military College has improved significantly since the introduction of the form. Unfortunately I cannot because data to support such a statement are not available.

Use of the form is voluntary and those who use it do so for their own information. I have asked members of faculty to pass their card decks to me, anonymously, so I could run an aggregate deck and publish college norms for each of the 22 critical requirements. I have no information which permits me to trace the progress of a single instructor over several administrations.[2]

---

[2] Except my own. Modesty prevents me from reporting how much I improved from year to year.

Fear keeps me from recording how low even the improved scores are.

Norms for the data which I have been able to collect are given in Table 2. Fortunately, it was possible to keep cards from four different sources separate and norms for psychology, management, one arts department and one engineering department, are shown.

Insert Table 2 about here

One study using the form was done by Bain (1973) in which he compares ratings of two professors teaching the same course, one with the traditional lecture method, following a very tight course outline, the other using a student centred approach with options for individuals to pursue personal interest topics and participate in group exercises.

The results are shown in Table 3.

Insert Table 3 about here

Of most interest is the result that students in the two groups did not assess one course (critical requirements 16 to 22) higher than the other, but showed enormous preference for the professor teaching the student centred group (critical requiements 1 - 15). The fact that they preferred one professor, or his teaching method, but did not overassess the course itself speaks for the usefulness and independence of the two parts of the evaluation form.

In summary, the RMC Professor and Course Critique Form allows students to assess professors and courses on critical requirements that the students themselves provided through Flanagan's C.I. technique. The main purpose of the form is to provide feedback to the professor on those areas of his teaching style and course material that students at RMC deem to be critical in good classes.

Many professors have said they like the form. In some departments all members use it, in other departments the distribution runs from few to none. Some professors are worried that results might get into the hands of Department Heads and Deans. Some press the administration to ensure that students do not rate professors and courses surreptitiously and publish student guide books.

The most common misunderstanding is characterized by the comment: "The items are not the ones I would have chosen - they are not the things I am worried about.".

With that quote I rest my case for the inappropriateness of off-the-shelf or professor constructed forms. Although the teacher may be the authority on content, only the student can tell him how he puts it across.

REFERENCES

Holmes, D.S. The Teaching Assessment Blank: A Form for the Student Assessment of College Instructors. The Journal of Experimental Education, 1971, 39, 34-38.

Mueller, R.H., Roach, P.J., and Malone, J.A. College Students' Views of the Characteristics of an "Ideal" Professor. Psychology in the Schools, 1971, 8, 161-167.

Gadzella, B.M. College Students' Views and Ratings of an Ideal Professor. College and University, 1968, 44, 89-96.

Flanagan, J.C. The Critical Incident Technique. Psychological Bulletin, 1954, 51, 327-358.

Otke, P.G. The Application of the Critical Incident Technique to the Study of Motivation and Leadership. Proceedings, 1969. 11th Annual Conference, Military Testing Association, 450-460.

Bain, E.H. Learning for Oneself and Interpersonal Growth Within the College Classroom. Report 73-2, Department of Military Leadership and Management, Royal Military College, Kingston, Ontario K7L 2W3.

TABLE 1

PROFESSOR AND COURSE CRITIQUE
CRITICAL REQUIREMENTS 1 TO 15: PROFESSOR, 16 TO 22: COURSE

No.	Rank		No.	Rank	
1.	(10)	Knowledge of subject	12.	(2)	Encouraged discussion
2.	(1)	Enthusiasm	13.	(8)	Class atmosphere
3.	(6)	Flexibility of schedule	14.	(9)	Pace of covering material
4.	(15)	Use of teaching aids	15.	(14)	Quality of explanations
5.	(12)	Lecturing ability	16.	(1)	Interest of course material
6.	(5)	Teaching effectiveness	17.	(4)	Quality of textbook
7.	(3)	Showed utility of material	18.	(3)	Challenge
8.	(7)	Showed interest in students	19.	(2)	Relevancy to future
9.	(4)	Well-organized	20.	(5)	Relationship to major
10.	(11)	Available for help	21.	(6)	Practicality of material
11.	(13)	Taught at students' level	22.	(7)	Amount of material covered

All CRs are scored from 1 to 5 with 1 being low, poor, etc., 3 is average or satisfies the requirement, and 5 is high, exceptional, excellent, etc. 2 and 4 are interpolated between the adjacent anchor points.

TABLE 2

AVERAGE SCORES ON THE TWENTY-TWO CRITICAL REQUIREMENTS BY DEPARTMENT

		Leadership and Mgmt Dept				
		Psych	Mgmt	One Arts Dept	One Engring Dept	Weighted Average
1.	Knowledge	3.88	3.96	3.90	3.86	3.91
2.	Enthusiasm	3.76	4.00	3.98	3.62	3.85
3.	Flexibility	3.50	3.65	4.00	3.45	3.61
4.	Teaching aids	3.18	3.17	3.63	2.70	3.14
5.	Lectures	2.92	3.37	3.67	2.89	3.18
6.	Teaching Effectiveness	3.23	3.40	3.56	2.96	3.28
7.	Showed utility	3.57	3.64	3.63	3.17	3.52
8.	Interested in students	4.12	4.28	3.78	3.63	4.02
9.	Well-organized	4.23	4.26	4.46	3.44	4.11
10.	Available	4.11	4.21	4.02	3.91	4.09
11.	At student's level	4.05	4.16	4.57	3.33	4.01
12.	Discussion	4.03	4.32	3.78	3.14	3.91
13.	Atmosphere	4.02	4.25	4.22	3.79	4.08
14.	Pace	3.26	3.05	3.79	2.89	3.18
15.	Explanations	3.40	3.45	3.49	3.04	3.36
16.	Interest of course material	3.60	3.56	4.02	3.19	3.56
17.	Textbook	3.47	3.34	3.63	3.13	3.38
18.	Challenge	3.18	3.06	2.43	3.43	3.08
19.	Relevancy	3.66	3.94	3.75	3.27	3.69
20.	Relationship	2.17	2.35	2.39	2.75	2.39
21.	Practicality	3.00	2.97	3.51	2.74	3.00
22.	Amount covered	3.33	2.88	3.73	2.90	3.14
	Departmental Average	3.53	3.60	3.72	3.24	3.52
	N	133	156	63	94	446

TABLE 3

TESTS OF SIGNIFICANCE OF DIFFERENCE BETWEEN MEANS FOR EXPERIMENTAL AND CONTROL GROUPS ON PROFESSOR AND COURSE CRITIQUE RESULTS

Critical Requirement	Experimental (N = 24) Group means	Control (N = 20) Group means	t or $\sqrt{F}$
1. Professor's knowledge	4.33	3.50	5.71 ***
2. Professor's enthusiasm	4.67	3.55	7.45 ***
3. Professor's flexibility... work schedule	4.67	2.55	8.18 ***
4. Teaching aids, use of	3.50	3.10	1.83
5. Lecturing ability of professor	3.04	2.50	2.89 **
6. Teaching effectiveness of professor	3.75	2.75	4.75 ***
7. Utility of course	3.67	3.40	1.27
8. Interest shown in students	4.67	3.65	6.31 ***
9. Organization and preparation by professor	4.45	4.35	0.61
10. Availability for help	4.12	4.05	0.35
11. Teaching at student's level	3.92	3.95	0.16
12. Encouraging discussion	4.75	3.70	5.19 ***
13. Class atmosphere	4.58	3.20	5.15 ***
14. Pace of material covered	3.96	3.25	3.10 **
15. Professor's explanations	3.63	3.20	2.54 *
16. Course interest	3.75	3.55	0.82
17. Textbooks	3.63	3.70	0.28
18. Course difficulty	3.42	3.05	1.52
19. Courses' relevancy to future	3.83	3.65	0.65
20. Course's application to your major	2.17	2.20	0.09
21. Theoretical content	3.17	3.10	0.25
22. Amount of material covered	3.88	3.50	1.73
Average rating for all CR's	3.90	3.34	1.73

Note: $^{*}p < .05$
$^{**}p < .01$
$^{***}p < .001$
Two tailed t-tests

PROBLEM CHECKLISTS, A PRAGMATIC APPROACH FOR FOSTERING UNIT EFFECTIVENESS

Frederick W. Timmerman, Jr., Ph.D.

United States Military Academy

Unit effectiveness is a goal that every commander works to achieve as he goes about the business of command. Usually, through measures such as AWOL rates, scores on AGI's, and CMMI inspections his superiors as well as the commander himself seek confirmation of a successful effort. The focus is on group performance outcomes. If they are good, as indicated by quantitative measures, the commander has reason to feel he is achieving the goal of unit effectiveness. If they are not, he is less certain what the unit problem might actually be. The quanitative measures certainly tell him what is wrong with the outcome behaviors but they don't always tell him why they are that way. In short, they don't give the commander the diagnostic tools to get at the "people" problems that certainly may underlie the poor performance outcome.

Faced with this dilemma, it would measurably assist the Commander if some behavioral tool were available that would: (1) allow identification of developing personnel problems before they were reflected in unit performance outcomes; (2) allow the commander the opportunity to correct the problem before it adversely effected unit performance. Aside from a very probable increase in command effectiveness, a reduction in discipline problems could certainly result from use of such a tool.

It is obvious, however, to be effective any behavioral tool would have to be aimed at identifying potential rather than actual soldier problems. This fact leads directly to the question of whether or not it is possible to predict which soldiers will indeed become problems for the commander. Berbiglia (1971), using the Taylor-Johnson Temperament Analysis personality inventory developed a predictive profile that was effective in identifying potential AWOL's. Nelson (1971) suggested that to predict broadly both job proficiency and emotional adjustment a combination of cognitive test data and biographical information was likely to provide the best index of soldier success at the early stages of his career. Fitts (1965b) remarked that the individual's concept of himself has been demonstrated to be quite influential in much of his behavior and directly related to his general adjustment state. Further, Mooney and Gordon (1950) remarked that problem check lists were effective for appraising the major concerns of a group and for bringing into the open the problems of each individual in the group. Finally, Cattell (1973) held that a general approach to assessing a person's adjustment state permitted one to recognize and handle a person's uniqueness as well as his common humanity.

As part of a larger study designed to determine if it were possible to predict which soldiers would become problems in line combat units (Timmerman, 1973), a comparative analysis of three personality inventories was undertaken. The analysis was designed to answer the following questions:

1. Was it possible using personality inventories and DA Form 20 data to predict which soldiers would experience adjustment problems?

2. Which inventory coupled with DA Form 20 data would do the best job of predicting the soldiers that would experience adjustment problems if prediction was possible?

Method

Subjects

The subjects were first term enlisted men with combat job classifications from the 82d Airborne Division at Fort Bragg, North Carolina. From a pool of first term soldiers designated as discipline problems (designation based on performance and disciplinary records) individuals were randomly selected for inclusion in the study. Non-discipline problem soldiers were then randomly selected from the remaining first term combat soldier population. Reduction in sample size due to missing records and cases where the data were incomplete resulted in a final available study sample of 193 subjects -- 73 discipline problem soldiers and 120 non-discipline problem soldiers.

Data Collection and Instruments

For the general study, data collection involved the gathering of information concerning the background and intellective capabilities of all subjects as reflected on the soldier's DA Form 20. For the comparative analysis the subjects completed three personality inventories. They included the Sixteen Personality Factor Questionnaire, (Form E) (Cattell, 1967), Tennessee Self Concept Scale, (Fitts, 1965a), and an author developed problem check list called the Soldier Problem Checklist. The Soldier Problem Checklist was patterned after the Mooney Problem Check Lists (Mooney and Gordon, 1950). It was developed in the following manner:

1. Potential soldier problem areas were identified through interviews with commanders and soldiers. These areas served as categories under which similar problems could be grouped and included:

   a. Health
   b. Economics
   c. Life location
   d. Self concerns
   e. Family
   f. Job
   g. Dating

2. Statements of problems that in any way might affect the soldier were collected and classified under the general headings.

3. Problem statements were reworded in the simplist language possible and repetitious problem statements were consolidated or discarded to keep the time for administration to about 30 minutes.

4. Two free response questions were added to develop additional data.

5. Scoring procedures and test directions were developed giving 16 scales for analysis purposes.

Procedure

In order to provide a means of cross-validating any predictive capability noted during the comparative analysis, the discipline problem soldier and the non-discipline problem soldier groups were further randomly divided into modeling and validation groups. A discriminant function analysis was then completed for each of the personality inventories administered and combined with the DA Form 20 data. Modeling groups data was used in the attempt to generate predictive profiles. Finally, cross-validation was accomplished using the independent validation groups.

Results

Table 1, Table 2, and Table 3 show the maximized classification matrices based on the discriminant analyses completed. Percentages of correct classification and the number of variables used, are shown for all prediction attempts.

TABLE 1

Classification Matrix Derived from Stepwise Multiple Discriminant Function Analysis Using Seven Predictor Variables from the 16 PF and DA Form 20 and Two Criterion Groups with Prediction to an Independent Sample

Group	Number of Cases Classified into Group		
	Problem	Non-Problem	Percent Correct
Problem	38	13	74.51
Non-Problem	19	62	76.54
Validation Problem	14	8	63.64
Validation Non-Problem	14	25	64.10

TABLE 2

Classification Matrix Derived from Stepwise Multiple Discriminant Function Analysis Using Eight Predictor Variables from the *Tennessee Self Concept Scale* and DA Form 20 and Two Criterion Groups with Prediction to an Independent Sample

Group	Number of Cases Classified into Group		
	Problem	Non-Problem	Percent Correct
Problem	35	16	68.63
Non-Problem	23	58	71.60
Validation Problem	13	9	59.09
Validation Non-Problem	20	19	48.72

TABLE 3

Classification Matrix Derived from Stepwise Multiple Discriminant Function Analysis Using Ten Predictor Variables from the *Soldier Problem Checklist* and DA Form 20 and Two Criterion Groups with Prediction to an Independent Sample

Group	Number of Cases Classified into Group		
	Problem	Non-Problem	Percent Correct
Problem	38	13	74.51
Non-Problem	18	63	77.78
Validation Problem	14	8	63.64
Validation Non-Problem	14	25	64.10

A comparison of Table 1, Table 2, and Table 3 revealed that the Soldier Problem Checklist paired with the DA Form 20 data using ten variables did a slightly better job of classifying cases than did the 16 PF (Form E) coupled with DA Form 20 data. Problem soldier modeling group cases were correctly classified 74.51 percent of the time, while validation group cases were correctly classified 63.64 percent of the time. Further, the pairing of the Soldier Problem Checklist and DA Form 20 data accounted for the largest percentage of modeling group variability for any of the three analyses. The pairing of the Tennessee Self Concept Scale and DA Form 20 data did not begin to match the other two combinations in classification capability.

Discussion

The results of the comparative analysis indicated that prediction was possible although only at a moderate level. The personality instrument that did the best job of predicting was the Soldier Problem Checklist. Based on the results and the additional fact that the Soldier Problem Checklist was originally developed with the aims of ease of administrating and scoring by a commander, it was selected for use in additional analysis.

In as much as the results showed the differential value of the personality instruments in prediction, they also indicated some pitfalls to be avoided. Personality inventories that assessed general or self adjustment rather than organizational adjustment appear to be less useful when assessing job adjustment. Moreover, we need not be sold on the idea that high powered commercial personality instruments are the only instruments that can do the job of assisting with personnel problems.

Any innovative or concerned commander should be able to develop a problem check list. If it reflected the operational reality of his environment it would be a useful management tool. When administered anomonously to the unit as a whole it could serve as a device to point out problems of importance to the group that might be hindering organizational efficiency. Additionally, use of the check list concept would bring to the commander's attention soldier problems that might have otherwise gone undetected. Prompt action by the commander could serve to foster improved relations between himself and his subordinates. In this way, problem check lists could be a pragmatic approach for fostering unit effectiveness.

References

Berbiglia, J.C. The AWOL Syndrome. Los Angeles: Psychological Publications, Inc. 1971.

Cattell, R.B. Personality Pinned Down. Psychology Today, 1973, 7 (7), 40-46.

Cattell, R.B. Sixteen Personality Factor Questionnaire, Form E. Champaign Ill: Institute for Personality and Ability Testing, 1967.

Fitts, W.H. Tennessee Self Concept Scale. Nashville, Tennessee: Counselor Recordings and Tests, 1965 (a).

Fitts, W.H. Manual for Tennessee Self Concept Scale. Nashville, Tennessee: Counselor Recordings and Tests, 1965 (b).

Mooney, R.L., and Gordon, L.V. Manual for the Mooney Problem Check Lists. New York: The Psychological Corp., 1950.

Nelson, P.D. Personnel Performance Prediction. In Roger W. Little (Ed.) Handbook of Military Institutions. Beverly Hills: Sage Publications, 1971.

Timmerman, F.W., Jr. Prediction of Enlisted Soldier Discipline Problems in Line Combat Units of the United States Army. Unpublished doctoral dissertation Duke University, 1973.

CLASSIFICATIONS OF ARMY OFFICER DUTY POSITIONS: IMPLICATIONS FOR TRAINING

by

Arthur C. F. Gilbert, Ph.D.

A Paper Prepared for Presentation at the
16th Annual Military Testing Association (MTA) Conference
October 21-25, 1974

Individual Training and Performance Research Laboratory
U.S. Army Research Institute for the Behavioral and Social Sciences
Arlington, Virginia 22209

CLASSIFICATION OF ARMY OFFICER DUTY POSITIONS: IMPLICATIONS FOR TRAINING REQUIREMENTS

Arthur C. F. Gilbert, Ph.D.

U.S. Army Research Institute for the Behavioral and Social Sciences[1]
Arlington, Virginia 22209

A portion of the research effort in the Army Research Institute for the Behavioral and Social Sciences (ARI) is concerned with development of methodologies to facilitate implementation of the Officer Personnel Management System (OPMS) (U.S. Department of the Army, 1974). One of the objectives of this research is to design a career progression model that can be used to reflect the skill and experience requirements in the different officer specialties at successively higher levels of responsibility in these specialties. This model will be used in the assignment of officers to those positions where their prior experiences are utilized to the maximum extent possible.

Design of an adequate career progression model requires describing in meaningful fashion the duties and responsibilities involved in the various duty positions that Army officers occupy. Furthermore, the system used in describing the duty positions must also responsively describe the experiences of the individual officer. One approach to this problem has been task analysis, but this has the definite limitation when used for the numerous diversified duty positions of military officers in that too many task statements need to be used to describe a particular duty position. For example, Lane and Marshall (1973) used 222 tasks statements in a study of the duties of Combat Information Center Officers in the Navy. Also, specific task statements are too fragmentary a vehicle for describing a broad spectrum of duties or for defining personnel or training requirements.

---

[1] The views expressed in this paper are those of the author and do not necessarily reflect the views of the Army Research Institite or the Department of the Army.

An approach that overcomes these limitations is the ''duty module'' concept. A ''duty module'' is a cluster of functionally related tasks, and a duty position may be described by one or more of these modules. The duty module rather than the specific task then becomes the basic unit for the description of a duty position. Not only does the duty module define the duties performed, but it can be used as a shorthand definition of the education or experiences of an individual. ARI has developed this duty module approach by contract with the American Institutes for Research and by in-house research.

In an earlier field study, Sitterson and Wintersteen (1974) developed duty modules in the Infantry and Quartermaster branches. Task-analysis information was collected on representative officer duty positions in these branches, and 93 duty modules were developed from this information. They then conducted a field survey, using these duty modules, of officers occupying duty positions in these two branches to determine the adequacy of the duty module approach in describing these positions under both combat conditions and garrison conditions. They concluded that the results of their field survey verified the adequacy of the duty module approach in describing officer duty positions in these two branches.

The purpose of this research was to determine the major underlying dimensions that describe Infantry and Quartermaster officer duty positions when these positions are described in duty module terms, using the data collected by Sitterson and Wintersteen. The extent to which these dimensions correspond to logical and functional groupings of Army officer duties provides an indication of the efficacy of the duty module approach. As a further evaluation of the duty module approach, the duty module data were analyzed for combat-related conditions and for garrison-related conditions to determine if the same dimensions could be used to describe both sets of conditions.

PROCEDURE

A total of 477 officers were used as subjects. Of these subjects 403 were in Infantry positions and 74 were in Quartermaster positions. Data were obtained that reflected the involvement of these officers in the 93 different duty modules in (1) actual or simulated combat operations and support, and (2) in garrison conditions.[2]

Separate parallel analyses were performed for the types of conditions (combat and garrison). Subjects were divided into groups on the basis of grade, position title, type of organization, and Military Occupational Specialty. For the purpose of these analyses, first and second lieutenants were pooled. Organizations were classified in terms of category (i.e., Airborne, Infantry, Mechanized Infantry or not combat deployable such as post-headquarters) and on the basis of function (e.g., headquarters, line, or support).

After the subjects were divided into groups, two data matrices were constructed. The rows of these matrices were defined by the groupings of subjects and the columns were defined by the 93 duty modules. The value of each cell entry in each matrix was defined by whether or not a particular group performed the duties described by a duty module. If a group performed the duties described by a particular duty module, then the cell entry defined by that group (row) and that duty module (column) was assigned a value of ‘‘1’’; if not, then the cell entry was assigned a value of ‘‘0’’. By this method two such data matrices were derived, one matrix representing combat and the other garrison conditions.

A matrix of the order 232 x 93 was derived for the data representing combat conditions; the matrix for garrison conditions was 312 x 93. For each matrix, those columns where less than five percent of the cell entries were ‘‘1’’ were then deleted. This process reduced the data matrix for combat conditions to 232 x 25 and the matrix for garrison conditions to 312 x 28.

An intercorrelation matrix was next computed from each of the data matrices. These correlation matrices were factor analyzed. A principal components solution was obtained where the highest absolute row value for each row was used as the diagonal estimate. Factors were extracted that had eigenvalues greater than 1.00, and these factors were rotated by the varimax method.

---

[2] Field collection of data was performed by American Institutes for Research under Contract DAHC-19-73-C-0042

RESULTS

Combat Operations and Support Conditions

Six factors with eigenvalues greater than 1.00 were extracted from the data on combat conditions. These factors accounted for 67.43 of the common variance. The percent of variance and the cumulative percent of variance accounted for by each of these factors is shown in Table 1. The percent of common variance and the cumulative percent of common variance accounted for by these six factors after orthogonal rotation is shown in Table 2.

Table 1

PERCENT OF VARIANCE ACCOUNTED FOR BY EACH OF THE SIX FACTORS IN ACTUAL OR SIMULATED COMBAT OPERATIONS AND SUPPORT ACTIVITIES

Factor	% Variance	Cumulative % Variance
1	19.73	19.73
2	12.75	32.48
3	11.95	44.43
4	9.65	54.08
5	7.95	62.03
6	5.40	67.43

The six factors have been defined as follows in terms of the magnitude of their corresponding eigenvalues in decending order: Factor I, Unit Command; Factor II, Operations and Training; Factor III, Manpower and Personnel; Factor IV, Logistics; Factor V, Intelligence; and Factor VI, Troop Welfare.

Factor I, Unit Command: This factor reflects a unit commander's activities in a variety of roles, ranging from unit administration to directing and controlling the tactical employment of troops, as indicated by the loadings of the variables on this factor (Table 3). Other variables that load substantially on this factor involve direct

Table 2

PERCENT OF VARIANCE ACCOUNTED FOR BY EACH OF THE SIX ROTATED FACTORS IN ACTUAL OR SIMULATED COMBAT OPERATIONS AND SUPPORT ACTIVITIES

Factor	% Variance	Cumulative % Variance
1	13.73	13.73
2	13.54	27.27
3	12.09	39.36
4	11.25	50.61
5	8.40	59.01
6	8.43	67.44

participation in combat, responsibility for supply operations in the support of the unit, the training of troops, and a general management function. The role of the unit commander in exercising his authority on judicial matters, which is solely the preogrative of the commander, is reflected in the duty module on the exercise of military command authority.

Factor II, Operations and Training: This factor reflects an operations and training constellation of functions at the staff level. As shown in Table 4 five variables have loadings in excess of .30 on this factor. These variables identify duty modules that involve general administration, operations staff functions, operations planning staff functions, training staff functions, and the organizing of staff functions. In the roles defined by this factor, the officer performs duties that lead to increased organizational effectiveness by providing advice on the organization of unit functions and the planning of manpower and material resources; these are reflected in the duty module on organization staff functions. Staff work directed toward defining the operational requirements of the organization and monitoring the execution of operational orders and directives is represented in the operations staff duty module, while the staff work involving the actual planning of operations is reflected in the operations planning module. Staff work essential to maintaining the operational readiness of the organization is reflected in the module on training staff functions.

Table 3

LOADINGS OF DUTY MODULE VARIABLES ON THE UNIT COMMAND FACTOR FOR COMBAT CONDITIONS AND FOR GARRISON CONDITIONS, AFTER ROTATION

Duty Module Variable	Loading	
	Combat Factor I_c	Garrison[a] Factor I_g
Performs unit administration	.77	.85
Directs and controls tactical employment of unit	.74	.49
Exercises military command authority	.73	.80
Trains troops and/or civilian employees in units and activities	.68[b]	.65[b]
Performs supply operations at consumer unit level	.67[b]	.77[b]
Participates individually and directly in ground combat	.63	--[c]
Performs command or general management	.36	.33
Supervises troop appearance and care and maintenance of material and facilities	.28	.39[b]
Counsels and evaluates subordinates as troop leader and takes action on personal problems	.25[b]	.38[b]
Supervises a staff section, detachment, or office	-.13[b]	-.31[b]
Performs general administration	-.29[b]	-.68

[a] Garrison loadings have been reflected.

[b] Variable has a substantial loading (i.e., over .30) on another factor in the same analyses.

[c] Variable not included in analysis of data for garrison conditions.

Table 4

LOADINGS OF DUTY MODULE VARIABLES ON THE OPERATIONS AND TRAINING FACTOR FOR COMBAT CONDITIONS AND FOR GARRISON CONDITIONS, AFTER ROTATION

Duty Module Variable	Loading	
	Combat Factor II_c	Garrison Factor III_g
Performs organization staff functions in a general staff or other coordinating staff	.92	.90
Performs operations staff functions in a general staff or other coordinating staff	.90	.89
Performs training staff functions	.89	.73
Performs operations planning staff functions in a general staff or other coordinating staff	.86	.82
Performs general administration	.37[a]	.19

[a] Variable has a substantial loading (i.e., over .30) on another factor in the same analysis.

Factor III, Manpower and Personnel: The four variables that load substantially on this factor in the combat-condition analysis are shown in Table 5. These variables represent duty modules that cover the supervision of a staff section, manpower management staff functions, and staff functions pertaining to personnel services. Generally, this factor reflects the planning and utilization of manpower resources as well as a personnel management function. The duty module on personnel services deals with the staff work necessary to provide for such matters as the spiritual welfare, housing and dining facilities, and the well-being of dependents, in order to insure the morale and well-being of personnel.

Factor IV, Logistics: This factor was defined by five variables. These variables, shown in Table 6, correspond to duty modules dealing with general administration; supervision of a staff section, detachment or office; performance of supply staff functions; the performance of

Table 5

LOADINGS OF DUTY MODULE VARIABLES ON THE MANPOWER AND PERSONNEL FACTOR FOR COMBAT CONDITIONS AND FOR GARRISON CONDITIONS, AFTER ROTATION

Duty Module Variable	Loading	
	Combat Factor III_c	Garrison Factor II_g
Performs personnel management staff functions	.98	.90
Performs staff functions pertaining to personnel services	.95	.91
Performs manpower management staff functions	.95	.84
Supervises a staff section, detachment, or office	.38[a]	.26[a]
Performs special staff administrative or adjutant type functions	--[b]	.74
Performs headquarters management staff functions	--[b]	.37

[a] Variable has a substantial loading (i.e., over .30) on another factor in the same analysis.

[b] Variable not included in analysis of data for combat conditions.

maintenance staff functions; and the performance of logistical staff functions. This factor reflects a group of functions concerned with the overall planning for and the acquisition of resources and material to facilitate unit operations as well as the preparation of estimates and reports on logistical matters. Staff work necessary for the more specific aspects of operational support is also involved, as well as responsibilities for general administration and the supervision of a staff section.

Table 6

LOADINGS OF DUTY MODULE VARIABLES ON THE LOGISTICS FACTOR FOR COMBAT CONDITIONS AND FOR GARRISON CONDITIONS, AFTER ROTATION

Duty Module Variable	Loading	
	Combat Factor IV_c	Garrison[a] Factor IV_g
Performs logistical services staff functions	.94	.88
Performs supply staff functions	.91	.85
Performs maintenance staff functions in a general staff or other coordinating staff	.90	.88
Supervises a staff section, detachment, or office	.38[b]	.30[b]
Performs general administration	.32[b]	.15

[a] Garrison loadings have been reflected.

[b] Variable has a substantial loading (i.e., over .30) on another factor in the same analysis.

Factor V, Intelligence: Four variables define this factor as shown in Table 7. These variables correspond to duty modules on general administration, supervision of a staff section, combat intelligence staff functions, and counterintelligence and security staff functions. This factor describes the functions of an intelligence staff officer who has responsibilities for planning production, preparation, and dissemination of combat intelligence data as well as responsibilities and duties for maintaining the internal security of the unit. General administrative duties and duties involving the supervision of a staff section are also related to the duty modules that define this factor.

Table 7

LOADINGS OF DUTY MODULE VARIABLES ON THE INTELLIGENCE FACTOR FOR COMBAT CONDITIONS AND FOR GARRISON CONDITIONS, AFTER ROTATION

Duty Module Variable	Loading	
	Combat Factor V_c	Garrison Factor V_g
Performs counterintelligence and security staff functions in a general staff or other coordinating staff	.96	.94
Performs combat intelligence staff functions	.95	.94
Supervises a staff section, detachment or office	.35[a]	.20
Performs general administration	.35[a]	.14

[a] Variable has a substantial loading (i.e., over .30) on another factor in the same analysis.

Factor VI, Troop Welfare: Four variables with substantial loadings on this factor, shown in Table 8, correspond to duty modules involving counseling and evaluation of subordinates, supervising troop appearance and care and maintenance of material and facilities in unit, performance of supply operations at the unit consumer level, and training. In general this factor is related to the duties that an officer performs in face-to-face relationships with troops at an operational level. A primary concern for their welfare appears in the module on counseling and evaluation of subordinates, while the module dealing with supervising troop appearance and maintenance of materiel expresses the responsibilities and duties of an officer concerned with an environment conducive to unit effectiveness and the comfort of his troops. Involvement in supply activities at the unit level also reflects a troop welfare activity, as does involvement in training responsibilities; however, these tasks appear to be less important than in the troop welfare factor in the unit command factor.

Table 8

LOADINGS OF DUTY MODULE VARIABLES ON THE TROOP WELFARE FACTOR FOR COMBAT CONDITIONS AND FOR GARRISON CONDITIONS, AFTER ROTATION

Duty Module Variable	Loading	
	Combat Factor VI_c	Garrison Factor VI_g
Supervises troop appearance and care and maintenance of materiel and facilities in unit	.91[a]	.82[a]
Counsels and evaluates subordinates as troop leader and takes action on personal problems	.90[a]	.84[a]
Performs supply operations at consumer unit level	.37[a]	.17[a]
Trains troops and/or civilian employees in units and activities	.31[a]	.23[a]

[a] Variable has a substantial loading (i.e., over .30) on another factor in the same analysis.

Garrison Conditions

In garrison-related activities, six factors with eigenvalues greater than 1.00 were also extracted. These factors accounted for 57.19 percent of the variance. In Table 9 the percent and cumulative percent of the variance accounted for by these factors is shown. The percent and cumulative percent of the common variance accounted for by these factors after orthogonal rotation is shown in Table 10.

The six factors that emerged showed broad similarity of content to those that emerged from the data on combat and combat-related activities. However, these factors differed in one respect in terms of the magnitude of their corresponding eigenvalues, in descending order which were: Factor I, Unit Command; Factor II, Manpower and Personnel; Factor III, Operations and Training; Factor IV, Logistics; Factor V, Intelligence; and Factor VI, Troop Welfare. It should be noted here that in the analysis of the garrison-condition data the sequence of magnitudes of the eigenvalues was slightly different from that in the analysis of combat-condition data. The factor reflecting manpower and personnel functions had a higher eigenvalue in the garrison-condition analysis than the factor reflecting operations and training functions, the reverse of this situation occurred in the combat-condition analysis.

Table 9

PERCENT OF VARIANCE ACCOUNTED FOR BY EACH
OF THE SIX FACTORS IN
GARRISON ACTIVITIES

Factor	% Variance	Cumulative % Variance
1	16.95	16.95
2	11.66	28.61
3	10.31	38.92
4	7.60	46.52
5	6.10	52.62
6	4.57	57.19

Table 10

PERCENT OF VARIANCE ACCOUNTED FOR BY EACH
OF THE SIX ROTATED FACTORS IN
GARRISON ACTIVITIES

Factor	% Variance	Cumulative % Variance
1	13.64	13.64
2	11.55	25.15
3	10.63	35.82
4	8.77	44.59
5	6.69	51.28
6	5.91	57.19

Factor I, Unit Command: As in the analysis of combat-condition data, this factor was defined by duty modules that ranged from unit administration to tactical employment of troops. The loadings of the variables on this factor are shown in Table 3 after being reflected. In this analysis, this factor tended toward bipolarity with two of the ten variables showing substantial negative loadings. The variable with the highest positive loading is unit administration, as in the combat-condition analysis, while the variable with the highest negative loading is general administration. Thus, one end of this bipolar factor represents the unit command activity while the other represents an administrative type of function. All of the six variables that had substantial loadings on this factor in the combat-condition analysis also had substantial loadings in this analysis. Two other variables that represented duty modules dealing with the counseling and evaluating of subordinates and with supervising troop appearance had substantial loadings on the unit command end of this bipolar factor. Also, two variables which did not define this factor in the combat-condition analysis had substantial loadings at the administrative end of this factor. These two variables represented duty modules on general administration and on the supervision of a staff section, detachment, or office. Of course, the variable that reflected participation in ground combat was not used in the garrison-condition analysis. Comparison of the loadings of the variables that had substantial loadings on this unit command factor reveals that this factor in the garrison-condition analysis is essentially the same as the first factor that was extracted in the combat-condition analysis.

Factor II, Manpower and Personnel: Five variables had substantial loadings on this factor, which was Factor III in the combat-condition analysis. The loadings for these variables on this factor are shown in Table 5. Three of the variables that had substantial loadings in the combat-condition analysis also had substantial loadings on this factor in this analysis: manpower management, personnel services management, and personnel management at the staff level. Two new variables not included in the combat-condition analysis (because they did not meet the criterion of at least 5% of the categories of subjects reporting involvement in these modules in combat conditions) also had substantial loadings on this factor. These variables represent staff management functions at the headquarters level and administration of a special staff. It should also be noted that the variable representing supervision of a staff section had a loading of .26 in the garrison-condition analysis.

Factor III, Operations and Training: This factor emerged with substantial loadings on the same four variables as in the combat-condition analysis, in which it was Factor II. Comparison of these loadings (Table 4) shows that the loading of the variable representing general administration was .19 in this analysis while it was .37 in the previous one.

Factor IV, Logistics: The fourth factor was again defined by substantial loadings on the variables that defined staff officer functions in the general field of logistics, supply, and maintenance (Table 6). The variable that represented the supervision of a staff section also had an appreciably high loading on this factor.

Factor V, Intelligence: Again clearly defined by this factor is the function of the intelligence staff officer, as indicated by the loadings of the two variables that indicate activities in the areas of combat intelligence and counterintelligence (Table 7).

Factor VI, Troop Welfare: The sixth factor again reflected a troop welfare function. This factor is defined by substantial loadings, as shown in Table 8, on the two variables that represent the counseling and evaluation of subordinates and the supervision of troop appearance.

DISCUSSION

The results of this research revealed that the duty positions held by Army officers in both combat-related and in garrison-related activities can be divided into six groupings or factors. These six factors describe the following types of function: unit command, operations and training, manpower and personnel, logistics, intelligence, and troop welfare.

The unit command factor is represented by a diversified group of duties centering on exercise of command functions in all areas of command responsibility; these define the unit command function as the execution of policies and procedures developed in staff activities. The functional staff factors reflect the specialized requirements of their respective areas of operations, personnel, logistics, and intelligence. The troop welfare factor appears to define the duties of an officer involved in face-to-face leadership of troops who has first-line responsibility for their well-being and motivation.

These findings tend to support current military eductional practices of providing the potential commander with a broad knowledge of staff functioning and capabilities as well as the proper application of completed staff work. Also, these findings tend to support current practices of providing highly specialized educational opportunities to those officers who are likely to occupy positions in staff agencies.

The data reduction strategy employed in this research by application of factor analysis appears to be a fruitful method for analysis of duty module data. Factor analysis permits a ready evaluation of the relative contribution of each duty module in the different constellations of functions that emerge.

Continued research on duty modules should lead to development of a system defining all Army officer duty positions in explicit functional terms. This system can then be used to (1) identify functional relationships among positions, (2) develop a career progression model, (3) relate training and experience to officer assignment alternative, and (4) relate training programs to assignment needs.

To elucidate, the relationships among duty positions can be clearly defined in precise terms by factor analysis. The underlying dimensions that define the revelant duty positions can be established, and the degree to which the duty modules are revelant to these dimensions can be evaluated. The overlap among duty modules will be evaluated to determine the degree to which the experiences gained in one duty module may contribute to performance in other modules.

The functional requirements of duty positions and the interrelationships among these positions, stated in duty module terms, will be used to define the career development patterns in the Army officer corps. The development patterns thus defined will be incorporated into a career progression model that will reflect the possible progression paths of the officer at successively higher levels of assignment in his military career. These patterns will also reflect the critical choice points in the officer's career, based upon the degree to which alternative assignment options are available in terms of Army requirements and experience in the different duty modules. Further refinements on the career progression model will incorporate not only the fact that experience in a particular duty module is required for a particular assignment but also the duration, recency, and quality of performance required in that duty module.

Assignment profiles for officers will be developed incorporating the assignment requirements defined by the career progression model. The assignment requirements can then be compared with available officers' assignment profiles. Selection for assignment will then become a matter of optimizing the congruency between the assignment profiles and position requirements.

In the same fashion that duty modules can be used to describe duty positions, duty modules can be used to describe the training requirements for different assignments. Again, as in the assignment paradigm, the experiences of the individual officer can be compared with the training requirements in the military course that he may be attending. This comparison will then determine the segments of that course which he is required to complete.

The career progression model will reveal at which grade or responsibility level there are significant additions or deletions of certain functions. Significant additions of new functions in a duty position or change in the scope of existing functions will then define the period in which further training or education is appropriate. Training may then be applied in a more timely fashion than when a potential incumbent completes the training requirements for a particular position a long time before this training is to be used.

REFERENCES

Lane, G. L., and C. T. Marshall. Occupational Analysis: Final Report on the Design of a Navy Officer Occupational Analysis System. WTR 73-31. Washington, D.C.: Naval Personnel Research and Development Laboratory, June 1973.

Sitterson, John D., Jr., and Joseph O. Wintersteen. Results of Field Survey to Evaluate an Experimental Set of Duty Modules. Silver Spring, Maryland: American Institutes for Research, January 1974. (Performed for ARI under contract)

U.S. Department of the Army. Officer Professional Development and Utilization. DA Pamphlet 600-3. Washington, D.C., 1974.

U. S. MERCHANT MARINE LICENSING EXAMINATIONS:
The Modernization of a Traditional Mission

by
Richard C. Willing, M.S.
Personnel Research Psychologist
and
David E. McDevitt, Lieutenant
United States Coast Guard Reserve
Staff
of the
UNITED STATES COAST GUARD INSTITUTE
Merchant Vessel Personnel Research, Development, and Testing Division

October 1974
A Presentation Made To
The Sixteenth Annual Conference of the Military Testing Association
Oklahoma City, Oklahoma

The United States Coast Guard is the only military agency of the United States of America concerned with the testing of civilians to qualify them to fill positions in a civilian industry. The U. S. Coast Guard is the issuing and testing agency for the federal licenses required of the officers of the United States Merchant Marine.

Regulation of the qualifications of the personnel of the U. S. merchant fleet began as far back as 1838. The Act of 7 July 1838 spoke to the manning of all steam vessels, passenger and freight, on navigable waters of the United States by competent personnel. It was hoped that this act would provide better security for the passengers aboard these vessels. The Act of 30 August 1852, commonly known as the Steamboat Act, was the foundation for the Steamboat Inspection Service. This act established Local Inspectors, appointed by the Collector of Customs, who would establish a licensing program for Engineers and Pilots.

A boiler explosion on the S.S. Sultana which killed 1700 people was the incentive for Congress to enact the Act of 28 February 1871 which founded the Steamboat Inspection Service. This act extended the licensing requirements for Masters and Chief Engineers and also provided the agency with the power to revoke the licenses which it had issued.

The Act of 30 June 1932 combined the Steamboat Inspection Service with the existing Bureau of Navigation to form the Bureau of Navigation and Steamboat Inspection. This nation-wide agency licensed merchant mariners but retained a localized testing system where each port was autonomous and testing standards varied greatly.

The infamous "Morro Castle" disaster brought forth the Act of 27 May 1936 which greatly expanded the requirements for fire protection on U. S. merchant vessels and was an incentive for increased stress of safety in licensing examinations. The Bureau of Navigation and Steamboat Inspection's responsibilities were increased and its name changed to the Bureau of Marine Inspection and Navigation.

The locally-graded and locally-administered essay examination was firmly established in the Bureau by this time, and this system remained in use through the transfer of the Bureau into the U. S. Coast Guard in 1942. Through the years certain sections of the examination such as Meteorology were re-written in multiple choice format. However, the largest portion of the examinations remained as essay questions.

In 1969 the U. S. Coast Guard contracted with the Educational Testing Service (ETS) for a research project designed to review the program used for examining individuals applying for licenses as deck and engineering officers in the U. S. Merchant Marine. The objectives of this program were to (1) identify shortcomings of the essay-oriented system then in operation and (2) if the situation warranted it, recommend modifications and or major changes which would result in the creation of a more valid examination program. The project was precipitated by complaints from the marine industry that the examinations were not valid indicators of a licensee's capabilities.

The examinations then in existence for licensed deck and engineering officers had subject matter areas specified by Coast Guard regulations, and the minimum number of questions on each subject was specified by U. S. Coast Guard Headquarters. All examining offices had identical card files from which to draw their complete examination which was then graded by local office personnel of highly varied background and experience.

In order to determine the actual functioning of this ostensibly structured examination program, staff members of ETS visited five Coast Guard licensing centers- New York City, San Francisco, New Orleans, Cleveland, and Miami- to observe and determine their operating procedures. To supplement this information, discussions were held with the Headquarters personnel responsible for preparing examinations and for overseeing the licensing operation. Additionally, U. S. Coast Guard files concerning this subject were examined to obtain further insights into the evolution of the examination system.

Discussions were then held with representatives of the United States Merchant Marine Academy at Kings Point, New York, and with officials at the State University of New York, Maritime College, at Fort Schuyler, New York. These discussions provided ETS with an understanding of how Merchant Marine officers were being prepared at institutions of this type; of how the staff members involved felt about the examining procedure; and of what changes they would like to see in the makeup of the examination.

Information and opinions from representatives of three large marine labor organizations actively involved in the training of deck and engineering officers were also obtained at this time. These organizations were the International Brotherhood of Masters, Mates, and Pilots; the Marine Engineers Beneficial Association, District I; and the Marine Engineers Beneficial Association, District II.

Two major changes regarding the licensing examination program were the result of these extensive discussions and investigations by ETS. At the present time, it is assumed that these changes will apply to all four levels of both the Engineer and Mate licenses. However, since the two lower level license examinations for the Mates and Engineers are the only ones which have been completed, the changes will only be discussed in reference to these. The changes were as follows: (1) Job specifications in regard to Second and Third Mate and Second and Third Assistant Engineer would be reevaluated. Because of changes in the Merchant Marine job structure since the Coast Guard took over licensing examinations in 1942, it was felt that this reevaluation was imperative in the construction of a valid examination program. (2) Essay questions would be eliminated from the Second and Third Mate and Second and Third Assistant Engineer examinations. There were five primary reasons for this decision: (a) the majority of personnel in the maritime industry preferred it; (b) The ETS study indicated that, due to the rotational duty policies of the U. S. Coast Guard, the expertise and experience of those people scoring the essay questions varied considerably. It was felt that eliminating this source of error variance from the examination procedures would markedly increase the validity of the examination program; (c) It would reduce the time necessary to take the examinations; (d) The Second and Third Mate and Second and Third Assistant Engineer examinations would not be scored in the examination room of the local office immediately after the completion of the examination. It was felt that eliminating this possible source of subjectivity from the scoring procedure would further reduce the error variance inherent in any examination program which is not completely standardized. Consequently, at this time the examinations are being scored at the U. S. Coast Guard Institute, Merchant Vessel Personnel Research, Development and Testing Division in Oklahoma City, Oklahoma. (e) The examinations would be prepared with the assistance of the maritime industry, maritime unions, and maritime educators. It is obvious that those individuals most involved in the industrial field can best determine what subject matter is revelant to the examination. Therefore, the content validity of the examinations would be maximized by using these experts in the process of test construction.

The process of investigating and establishing relevant job specifications for the licensed positions of Second and Third Mate and Second and Third Assistant Engineer was difficult and time consuming. In order to accomplish this task, consultants were retained who wrote two or more job specifications and met in meeting and discussion groups over a period of many months to resolve differences in their written specifications. These consultants in the specification development were volunteers from many marine concerns: American President Lines; The Brotherhood of Marine Officers; Cleveland-Cliffs Iron Company; The State University of New York, Maritime College; Gulf Oil Corporation; Humble Oil Company; Interlake Steamship Company; International Organization of Masters, Mates, and Pilots; The U. S. Merchant Marine Academy at Kings Point, New York; Lake Carriers Association; Marine Transport Lines; Marine Engineers Beneficial Association, District I; Marine Engineers Beneficial Association, District II; Mobile Oil Corporation; Oglebay Norton Company; Prudential-Grace Lines; U. S. Coast Guard; U. S. Maritime Administration; U. S. Military Sealift Command; and U. S. Steel Corporation. With such a diverse group of interested parties involved in the process, it is no surprise that the ETS report indicates that great difficulty was encountered when an attempt was made to arrive on a consensus of specifications. Substantial acceptance was, however, eventually obtained. This assertion is supported by the fact that, of the complaints received about the new examinations, few have involved a disagreement with the subject matter areas which make up the complete examination.

After final Coast Guard approval of the job specifications for Second and Third Mate and Second and Third Assistant Engineer, contracts were let with various members of the maritime industry to acquire the raw test items which eventually, in refined form, would make up the licensing examinations. Item writers were obtained from the same source as the consultants for the job specifications, i. e., the maritime union training institutions, the national and state maritime academies, and the maritime industry. Item writers for the Second and Third Mate examinations held Master or First Mate licenses while those who wrote items for the Second and Third Assistant Engineer license examinations held Chief Engineer or First Assistant Engineer licenses.

U. S. Coast Guard personnel were not utilized in the writing of the original test items. Modifications of original items were, however, written by Coast Guard Officers and included in some forms of the final examinations. These few items did not make up more than two or three percent of any one examination.

The item writers sent their completed items to ETS headquarters in Princeton, New Jersey. After all the items were sent in, panels of maritime industry personnel were organized to review the items for clarity, conciseness, correctness, and relevancy. Many non-valid items were deleted. Final refinement and organization of the items into pre-test form was accomplished at ETS. Experimental pre-tests for Third Mates and Third Assistant Engineers were administered to graduating seniors at the U. S. Merchant Maritime Academy at Kings Point, New York, in the spring of 1972. Test analysis following this administration resulted in the elimination of additional items of questionable validity. For the remainder of 1972, ETS organized the total supply of available items into test forms deemed suitable for printing. This examination material was then sent to the newly organized Merchant Vessel Personnel Research, Development, and Testing Division of the U. S. Coast Guard Institute in Oklahoma City, Oklahoma, in early 1973 where the material underwent further extensive refinement. The final forms of these examinations were printed in late 1973.

The new standardized examinations were first administered in December, 1973 to the Texas A & M, Maritime Academy and have been administered on a monthly basis since then. After each administration, answer sheets from the U. S. C. G. Marine Inspection Offices throughout the country are forwarded to the U. S. Coast Guard Institute (MVP) in Oklahoma City where the examination undergoes analysis by an IBM 360 computer. Standard item analysis programs are applied in this procedure. When the computer printout designates certain items as being of questionable validity, they are evaluated by the division's U. S. Coast Guard Officers, Personnel Research Psychologist, and Testing/Training Specialist. The experience of the officers, all of whom hold licenses in the U. S. Merchant Marine, is immeasurable helpful during this review process.

Officers from the division make periodic field trips to the various maritime union training institutions and merchant marine academies, particularly those which have a large number of students who have recently taken the analyzed examination, to discuss matters relating to the examination. These discussions have brought forth many useful comments which have been incorporated into the examination. These comments have enabled the Coast Guard to modify portions of the examinations to increase their validity as job performance indicators. The U. S. Coast Guard feels this continual evaluation, reevaluation, and consequent modifications will result in a lower level licensing examination program which is unsurpassed in validity by any existing job performance examination program.

In the near future, similar procedures will result in a valid licensing program for Master, First Mate, First Assistant Engineer, and Chief Engineer licenses. The type of questions will vary somewhat from those used in the lower level examinations, but their validity should be equivalent to them if competent item writers can be employed. It is anticipated that there will be little difficulty in accomplishing this task. When the upper level examinations are constructed utilizing these items, the goal of an objective, standardized, and valid examination program for the U. S. Merchant Marine will be realized.

SECTION XVI

SYMPOSIUM:

EVALUATING JOB PROFICIENCY

FUNCTIONALLY INTEGRATED PERFORMANCE TESTING

by

W.C. Osborn, J.H. Harris, & J.P. Ford

Human Resources Research Organization

Paper prepared for:

Military Testing Association Conference
Oklahoma City, Oklahoma
October, 1974

FUNCTIONALLY INTEGRATED PERFORMANCE TESTING

The US Army has an undeniable need for valid job proficiency tests to support its personnel management system. These tests provide information vital to sound training and personnel decisions. Job proficiency tests play a critical role in decisions relating to such areas as: evaluating training, evaluating MOS proficiency, evaluating unit readiness, and as a criterion for estimating the effectiveness of experimental or administrative interventions in the personnel system.

Training and personnel managers typically use one of three general types of evaluation instruments to assess job proficiency: ratings of job performance, job knowledge tests, and job performance tests. Ratings -- whether contributed by supervisors, subordinates or peers -- represent the more convenient, inexpensive and widely used method of evaluating job proficiency. Unfortunately, they not only lack the standardization which tests have, but they are fraught with various rater errors and biases.

Job knowledge tests are also widely used. But there lingers a question concerning the degree to which these tests can adequately assess a person's job performance capability -- that is, both in the sense of the range of job performances that are validly represented in job knowledge, and in the sense that knowledge testing in a paper-and-pencil mode presumes at least minimal literacy skills.

Job performance tests are generally conceded to have the greater validity. They call for application of knowledge and demonstration of skill by eliciting behaviors that are equivalent, or nearly equivalent, to those required in actual job performance. However, because of excessive demands on time and resources, performance tests are generally considered to be too expensive.

Research is underway to strengthen these methods of job proficiency evaluation, and you will hear examples of this research described today.

The starting point in any program of research in criterion development is a maximally relevant criterion measure. That is, before we can explore the relative validity and efficiency of other more feasible measures, we need something against which more usable alternatives may be compared -- a near-ultimate measure of criterion performance.

The concept I wish to describe today grew out of the early planning for just such a project. Under sponsorship of the Army Research Institute, we are attempting to develop a maximally job-relevant performance test for a combat occupational specialty -- a test perhaps too costly to use for on-going proficiency evaluation programs, but one which can be used experimentally as a best-possible criterion for testing the relevance of other measures of job competency.

In our early thinking about this project we realized that typical job performance tests have a major shortcoming: they tend to leave out important aspects of a person's job requirements. This tendency to evaluate only part of a job incumbent's responsibilities is not inherent in performance tests; rather it reflects the current approach to developing performance tests. Where performance tests are used in the Army, mainly in Basic Combat Training and some Advanced Individual Training programs, they tend to focus on discrete job tasks. Typically, a test administrator orients a soldier to a job setting, tells him what task to perform and the time limitation, if there is one. After the soldier completes one task, the tester orients him to another task. The soldier thus progresses task-by-task until he has demonstrated mastery of the minimum job requirements. This approach precludes testing on some tasks and, for others, ignores a vital component of task mastery -- the soldier's ability to initiate his own performance from cues he would typically receive on the job. Logistical considerations have led to testing task-by-task, station-by-station. Thus, it is not surprising that important aspects of job performance are excluded. It would be of little benefit, for example, to inform a soldier, "At this station you will demonstrate your ability not to smoke around gasoline" or, perhaps a more subtle example would be, "You have just detected a noxious odor that you believe to be poison gas. Give the alarm." In both cases important aspects of task performance are excluded. Not smoking around gasoline is a capability that anyone has. The essence of the performance is some combination of attentiveness to the hazardous stimulus (gasoline) and a willingness to respond (not smoke). In the CBR example the critical performance is detecting the presence of gas when unalerted to that possibility. Actually giving the alarm, while important, is the more perfunctory element of task performance.

Shortcomings of the task-by-task approach to performance testing thus center around occasional failures to incorporate important aspects of task performance in the test situation. The main reason for this is that tasks of the sort mentioned are typically time shared and call for a response under unalerted conditions. Thus, any attempt to structure a separate test on one of these tasks necessarily preempts an unalerted set.

A possible solution to this problem lies in an approach to job performance testing which evaluates not only the soldier's mastery of skills and knowledges, but also his ability to recognize and react to conditions which initiate task performance. The tests evaluate mastery of each task in a job-related module which includes several other tasks. Because each task is tested in a functional context which maintains the logical relationship of the tasks, the approach is termed here "functionally integrated performance testing." This is to distinguish the approach from the currently typical approach of testing each task as a discrete incident.

The concept of functionally integrated testing is based on an analysis of task performance that indicates three types of stimuli which control initiation of task performance:

1. Leader
2. Environment
3. Self

The first type of stimulus is the command of a leader. A soldier is typically told when to begin performing such tasks as loading the main gun on a tank, classifying a bridge, or executing the manual of arms with a rifle. Tasks of this sort lend themselves well to conventional testing as separate and individual tasks.

The second type of initiating stimulus is controlled by the environment. It is worth distinguishing two varieties of environmentally controlled tasks. One variety involves immediate reaction to infrequent and unpredictable stimuli. A stimulus condition may be blatantly obvious, as a rifle misfiring which cues a reaction to clear the weapon, or the appearance of two or more enemy targets which cues a target selection decision. On the other hand, the stimulus condition may be subtle and require vigilance, as in avoiding contact with poison plants or detecting an enemy infiltrator. But in either case the stimulus is unpredictable if not totally unexpected, and an appropriate test situation should create an unalerted set on the part of the person being tested.

The third type of stimulus is from within the soldier himself. Tasks with a substantial affective component such as maintaining noise discipline, performing during-operations checks on a vehicle, and servicing an individual weapon typically fall into this category. For tasks in which the essence of the criterion behavior is affective, the question is not can the individual perform the task, rather it is will he perform the task in the presence of competing response tendencies. Obviously, in attempting to validly assess if the soldier will perform a task, some unobtrusive method of testing must be used.

As a demonstration of the difference between functionally integrated testing and discrete-task testing, consider the mission of occupying and operating an observation post. This mission encompasses six tasks among others:

1. Select a position for an individual observation post
2. Place the TA/1 telephone into operation
3. Camouflage field wire
4. Perform communications check
5. Observe sector of terrain
6. Transmit a spot report

A discrete-task job performance testing approach would result in evaluating mastery of each task separately. The soldier would be told which task to perform and when to begin. Further, Tasks 2, 4, and 6 would typically be tested at a communications station and Task 1 would be tested at a tactics station. Tasks 3 and 5 might very well not be tested at all because they seem to be relatively simple and non-proceduralized.

A functionally integrated job performance testing approach would group the tasks into a module. The orienting instructions would tell the soldier to set up an observation post, designate an observational sector, and require him to perform communications checks at a specified interval. Thus, the stimulus for performing Tasks 1 and 4 is initiated by the leader. Task 2 might appear to be environmentally initiated since the soldier could not perform a communications check without placing the telephone into operation, but it has neither the vigilance or immediate reaction features of an environmentally cued task, and as it is so closely tied to the communications check it is better viewed as a leader initiated task. Tasks 5 and 6 are controlled by the environment. Observing a sector of terrain requires sustained attention in order to detect the presence or activity of the enemy, and transmission of a spot report is initiated by and a reaction to that detection. Task 3 is initiated by the soldier himself. Camouflaging field wire does not require much skill of the soldier, but it does require recognition of the situation that calls for the task and a willingness to perform the task. As with many tasks that are essentially safety precautions, a personal commitment to camouflaging the wire is of particular significance because the system will operate perfectly well without it, therefore under pressure the soldier may overlook it or be tempted to ignore it.

Because it preserves the initiating cues of the real work environment, the functionally integrated test should be able to claim a higher level of validity than more conventional performance tests. That is, training or personnel managers should be more confident that a man who passes the functionally integrated test will perform his job acceptably than they would if the man passed the six discrete-task tests. Further, the higher validity may not raise the cost of the performance test, since the functionally integrated test would require no more equipment than would tests on the separate tasks. It is even possible that the functionally integrated test may be cheaper because it eliminates the time soldiers spend moving between stations, and testers spend reading orienting instructions. Of course, in economy, it cannot compare with group administered paper-and-pencil tests or with ratings of job proficiency, but it should at least offer a highly relevant criterion against which they can be objectively validated.

TEST FIDELITY: CONGRUENCE OF TEST CONDITIONS WITH ACTUAL ARMY TASKS

EUGENE JOHNSON, III AND ANGELO MIRABELLA

TEST FIDELITY: CONGRUENCE OF TEST CONDITIONS WITH ACTUAL ARMY TASKS

INTRODUCTION

The increasing movement towards performance based, hi-fidelity, full scope testing in the Army has come into conflict with an equally urgent requirement to reduce the cost and environmental impact of both training and testing, particularly for combat readiness.

The solution to this conflict is to develop--for appropriate job tasks--simulated approaches to testing--approaches which can meet the logistical constraints imposed by full-performance testing and yet provide acceptable levels of predictive validity.

In order to implement this solution, guidelines are needed for deciding when simulation is appropriate and how far the testing can deviate from operational conditions before the testing is hopelessly compromised as a means for making decisions about training, selection, or classification.

Currently, such guidelines are lacking, because almost no research into the problem of test fidelity has been done. Some work has been done in this area by the Enlisted Evaluation Center, the Army Research Institute, Matrix Corporation, and by HumRRO. But these have been primarily developmental efforts aimed at producing and evaluating specific tests. The picture which has emerged from them is not very clear and has not resulted in practical guidance on how and when to prepare simulated tests.

The purpose of this presentation will be to describe a general framework for looking at the test fidelity problem, to review the state-of-ignorance-of-the art, and to summarize a research effort which is currently underway at ARI.

If you have taken the Disney "World Tour", you have experienced high fidelity visual simulation--perhaps the best in the world available to the general public. Examples of high fidelity simulation abound in NASA's Apollo program, including the very difficult to accomplish zero gravity and lunar gravity simulation. An interesting example of fidelity comes to mind from an Army infantry training situation. In the late 1950's a life-sized, human-shaped target, which would fall to the ground when hit in a vital area, was developed by Human Resources Research Office. I believe it was called Punchy Pete by the troops. Two important aspects of fidelity were demonstrated by Punchy Pete. Firstly, even though it was only a cut-out (not very high on the physical fidelity dimension) the experience of firing at Pete was very different from firing at a bull's eye target. There was emotional impact. Pete was high on the dimension of psychological fidelity. Secondly, when you hit Pete, you knew it immediately, much as you would in a front line confrontation. That is high feedback fidelity, and you do not get the same effect with static paper targets.

It should be clear from the illustrative examples above that the concept of fidelity with which we are concerned does not necessarily depend upon total physical identity between two stimulus-response fields. Rather, we are focusing our research on psychological equivalence. More specifically, we are concerned with the principles involved in establishing training and evaluation situations which are psychologically or functionally equivalent to actual components of army tasks.

Psychological equivalence, like beauty, is in the eye of the beholder. This fact has been demonstrated many times by ethologists who study the stimuli which control animal behavior. For example, in the stickleback fish, mating behavior is initiated in the presence of small rounded objects with red markings. The presence of a mate is not required. The red-streaked, rounded object is a low physical fidelity stimulus with high psychological fidelity. Punchy Pete can be viewed in a similar context.

Our research program at the Army Research Institute is concerned with fidelity in two different, but highly related contexts. The first context is training. We are exploring the principles involved in designing training situations that will prepare our men to perform actual Army tasks safely and efficiently, where full field training is limited by constraints such as feasibility and cost. The goal is to maximize transfer of training between the training situation and the actual army task. High positive transfer requires high psychological fidelity between training and task. The second context is evaluation or testing. One goal of an evaluation program is to provide the basis for making inferences about how well men are prepared to perform assigned Army tasks. This goal requires valid tests, and valid tests require high psychological fidelity or equivalence between the actual Army task and the best situation.

The ideal training and evaluation program was approached in NASA's Apollo series. Needless to say, the program was very expensive, more expensive than today's Army training and evaluation budgets can provide. Our research program, then, is directed toward optimizing between fidelity and cost. Our goals are to achieve the greatest positive transfer of training and the most valid evaluation at the least possible cost. The key concept is meeting both of these goals to psychological fidelity.

An Analysis of Fidelity.

There are at least four important facets of situations to analyze when attempting to arrive at a global index of test-task fidelity. The first involves how behavior is initiated, evoked or emitted. The event(s) which antecedes behavior is generally called a cue or stimulus. The stimulus can arise from the external environment or it can be produced by the behaving organism. Bill Osbourn's paper which you have just heard dealt with this facet of fidelity. Recall that he talked about the need for the soldier to respond to "real" environmental cues without prompting, as well as to the soldier's own feelings about what he ought to do and wants to do. The second facet for analysis is the behavior itself. A unit of behavior is generally called a response. The first facet is the perception of the consequences of the response. It is often called response-produced feedback. It has been argued rather convincingly by some behavioral scientists that all behavior can be described as a seeking of appropriate response-produced feedback. For example, when a rifleman sees Punchy Pete (stimulus) pop up, he fires (response) until Pete falls down again (appropriate response-produced feedback). Although the stimulus, the response and the feedback in many Army tasks are not as clearly defined as in the Punchy Pete example, nevertheless, these three facets must be considered when analyzing test-task fidelity or equivalence. The fourth facet of a situation which often is overlooked in a fidelity analysis is ambient or task-peripheral stimulation. For example, intense background noise such as experienced in the midst of battle can have a great effect on marksmanship. So during an evaluation of marksmanship using Punchy Petes, it might be wise to include battle sounds (ambient stimulation) played over a loud speaker.

To summarize: Four important factors to analyze in determining test-task fidelity are (1) stimulus, (2) response (3) response-produced feedback, and (4) ambient stimulation.

In addition to the four factors mentioned above which are critical to a fidelity analysis, another factor, which we shall call scope, must be considered. Scope refers to variations which can be expected in the other four factors. For example, an infantryman could encounter an infinite variety of potential targets (stimuli), fire (response) at those targets with an infinite variety of weapons, perceive an infinite range of consequences (feedback) from his firing, and experience an infinite variation in noise level and weather conditions (ambient stimulation). But it is not feasible to train and to test that soldier ad infinitum (although some psychologists I know would feel more secure if that requirement came from a sponsor!) Since testing time is both expensive and limited, test fidelity can only approach perfection. Some sampling must occur from the vast variation possible in even the simplest situation. For example, a rifleman might be tested using only three Punchy Petes; one 45° left, one straight ahead, and one 90° right. Perhaps the test would be administered with the sun overhead and at dusk. Thus, three target positions and two lighting conditions are sampled from an infinite set of potential ones. The larger the sample of situations tested, the greater the test fidelity, the greater the certainty about the capabilities of the infantryman, and, in general, the greater the cost. So the cost of testing must be weighed against the consequences of sending men to the field whose performance capabilities are uncertain. This fact is all too well known to those of you who are directly involved in training and/or evaluation. And you also know that the Army has made a commitment to broad scope, high fidelity, performance testing. And you also know that the training and evaluation budget is more likely to remain constant or decrease, than it is to increase.

The way out of this dilemma is demanding, tricky and psychologically fascinating. It is demanding because it requires a broad and deep knowledge of the experimental literature on human performance. How important is variable x? Can stimulus a be substituted for stimulus b? If one can perform x can he be expected to perform y without additional training? It is tricky because the answers to the preceding questions always seem to lie in experiments yet undone, and the Army frequently cannot afford to wait for those experiments to be done, so the answers to those questions either come from good intuition plus knowledge or from the Army itself through trial and error. The fidelity-cost dilemma is psychologically fascinating because one way to reduce the cost of evaluation is to use imagination--both the test constructor's imagination and the test taker's imagination.

An imaginative test constructor will try to find psychologically equivalent but less costly stimuli, response, feedback and ambient stimuli from which to build tests. Initially, he may develop a full scope, high fidelity, performance-based test, requiring extensive time and resources to administer. You have heard Bill Osbourn discuss this notion. It is a notion which we at ARI are currently using in a unit training context to develop a test bed, particularly for tank gunnery training. In addition, he may construct a test which requires only paper and a pencil to administer. However, in general, the most useful test--one which is lower in cost than the full scope performance test and more valid than the paper and pencil test--will require the imaginative use of the principles of psychological equivalence.

Imagination will be required of the test taker as well to facilitate the realism of tests using psychologically equivalent elements. For example, when a pilot trainee is being tested in an aircraft simulator, the test probably will be more valid if he thinks, "I am flying an aircraft", than if he thinks, "Oh, it's just a simulator". One of the most fascinating characteristics of the human brain is its power to construct a total reality from bits and pieces of reality. Test constructors must learn more about how to capitalize on this power. Magicians have been doing it for centuries.

Measuring Psychological Equivalence or Fidelity

Psychological equivalence between test-task components can be measured in basically two ways. The first is by some judgmental or scaling process which might be called a consensus of experts. Extrapolations can be made from existing data to the specific situation which is being analyzed. Judgments from a set of experts can be combined in an informal way through discussion and eventual agreement, or formalized psychometric procedures may be utilized. One of the Army Research Institute's contractors, American Institutes for Research, has developed a formal psychometric model which yields a fidelity index. Elements or components derived from task analyses of the situations, i.e., jobs, tasks, training programs, tests, being compared are ratings for similarity by experts, and the ratings are combined into a global index of fidelity. Psychological equivalence as well as physical fidelity is reflected in the index.

The second way to measure psychological equivalence is through experiments designed to answer specific questions. This route considers psychological equivalence between test and task to be an empirical question. This approach is both more costly and more certain.

The two approaches to measuring fidelity are intimately related to one another. The stronger the data base derived from the more costly empirical method, in the long run the greater the probability of accuracy using the less costly consensus of experts method.

Application of the Concept of Fidelity: Ongoing Research.

Several Army Research Institute research projects heavily depend upon the analysis and measurement of fidelity. For example, Matrix Corporation has developed a set of tests called symbolic substitutes which only require paper and pencil. These tests make extensive use of photographs, diagrams, and descriptive scenarios to provide psychological equivalents of task stimuli. An electronics test, for example, requires the man being tested to draw pencil lines on a diagram to illustrate how he would wire a piece of equipment. An armor test requires the tank driver under evaluation to trace a potential course on a terrain map. A thorough evaluation of the validity of these tests has not been completed yet, but it is likely that they will have higher validity than verbal tests of knowledge about these tasks.

Human Resources Research Office has begun work on a testing concept called synthetic testing. Synthetic test construction requires an exhaustive analysis of task substructure. Then critical components of the task are sampled and part-task simulation techniques are used to develop test materials. A task evaluation may not require the performance of an entire task, but different forms of the test may require performance of any of several components of the task. Thus, the trainee must be prepared to perform the complete task.

It is through research efforts like the two examples mentioned above that the Army Research Institute is trying to help the Army more toward valid performance testing while holding costs down. The key concept around which this program revolves is psychological fidelity. In both the training and the evaluation contexts we must learn what components are necessary and sufficient to create situations which are psychologically equivalent to actual Army tasks.

DEVELOPMENT OF CRITERIA RATINGS TO MEASURE JOB PERFORMANCE*

by

Clifford P. Hahn
American Institutes for Research

INTRODUCTION

This report is a follow-up on the study presented initially at last year's Military Testing Association meeting at San Antonio. The Air Force sponsored study is designed to develop performance criteria at the job task level. A distinguishing characteristic of the study is the use of 50 or more task-level performance indices per job as opposed to a few more global measures.

Data were collected on task level performance of airmen at the journeyman level in three different career fields. All of the data reported on in this paper is from one of those fields, namely the Communication Center Operations career field. Performance data were collected at the task level from job incumbents, their supervisors and peers. In addition to reporting ratings at the task level, supervisors and peers also supplied ratings on six general ratings. All of the data were collected through mail administration which was channeled through a career field enlisted coordinator at each base.

The official Air Force Job Inventory for the Communication Center Operations field contained a total of 271 separate tasks. Through a series of iterative procedures using information about the number of incumbents in the career field who performed the task as well as the relative amount of time spent on the tasks and the expertise of experienced NCOs in the field, this list was reduced to 44 tasks which were judged to be the most critical. Critical in the sense that performance on them discriminated between especially effective and less effective airmen. For five of these tasks two dimensions and for one additional task, three dimensions of performance were judged to be critical. Thus a total of 51 task level performance dimensions were rated.

In addition to the task level data, supervisors and peers were also requested to supply six general performance ratings on percentile scales. One was an overall total job performance scale. Two of the others were what might be called product-oriented scales, i.e., amount of work produced and quality of the work produced. The remaining three were more motivationally mediated, i.e., willingness to do more than his share of the work; willingness to start and continue work on his own without the need for prodding, and willingness to share his skill and knowledge with peers and subordinates.

---

*Paper presented as part of a Symposium on Evaluating Job Proficiency at the 1974 Military Testing Association Convention, Oklahoma City, Oklahoma, 21-25 October 1974.

ATTRITION

Task level performance measurement necessarily involves a large amount of effort per case since 50 or more indices are collected rather than one or a relatively few. If one wishes to study individual tasks, the question arises how many overall cases are needed in order to ensure adequate numbers at the individual task level. Our initial survey materials were mailed to approximately 2000 cases. From this we received usable performance ratings for 1034 cases. Thus even after repeated mail and telephone follow-up, we lost approximately half of our sample through non-returns. Our feeling in retrospect is that had we utilized the official survey officer at each base instead of volunteered enlisted coordinators from the career field itself, the basic return rate would have been higher.

Obviously not all raters responded to all task dimensions listed. Incumbents were asked to respond to those which they felt applied to their specific duty position. Supervisors and peers were asked to respond to those which they felt were part of the incumbent's duty position and the performance of which they were able to observe. The "Not Rated" column in Table 1 shows the amount of attrition that occurred at the task dimension level. All task dimensions were rated by at least 20 percent of the total sample of incumbents, but none was rated by more than 70 percent. Only 29 of the 51 dimensions were rated by over half of the incumbents.

Similar data for supervisor ratings are shown in Table 2. All task dimensions were rated by at least 10 percent of the total sample of supervisors, but none were rated by more than 70 percent. Only 20 of the 51 dimensions were rated by over half of the supervisors. Table 3 shows similar data for peer ratings. One task dimension was rated by less than 10 percent of the total sample of peers and none was rated by more than half of the sample of peers.

The incumbents were most inclusive, the peers least, and the supervisors somewhere between which might be expected. The point to be remembered is that when you move from global ratings to specific task ratings, the attrition is great. The most frequently rated task dimension by incumbents was rated by 696 of 1034 incumbents and the least rated was rated by only 246. For supervisors equivalent frequencies were 664 and 114 and for peers 452 and 90.

RATING DISTRIBUTIONS

One of the questions the study addressed is whether or not a task level performance approach could successfully capture more of the performance variance than global ratings usually did and whether or not task level ratings could ameliorate the strong skew of existing general ratings. The distributional characteristics shown in Tables 1, 2, 3 and 4 would seem to indicate that the problems of restricted variance and strong skew were not completely overcome by the task level approach.

A review of Table 4 indicates that incumbents tended to rate themselves somewhat lower than did their supervisors and peers, and the variance in their ratings was somewhat higher. Incumbent mean ratings ranged between 4.97 and 6.35; supervisor means ranged between 5.73 and 6.66; and peer means ranged between 5.72 and 6.60. The variance of incumbent ratings ranged between 1.00 and 1.80; for supervisors between .85 and 1.60; and for peers between .88 and 1.56. For all separate task dimensions, the mean incumbent rating was lower than that from supervisors or peers. For 26 of the dimensions the supervisor means were highest, for 2 dimensions supervisor and peer means were equal, and for the remaining 23 dimensions the peer means were the highest. For all six of the general ratings the supervisor means were higher than peers and the variance of supervisor ratings were all less than of peers.

All of the task dimension ratings were decidedly positively skewed toward the high performance end of the 7 point scale used. A review of Table 1 indicates that incumbents used all of the rating points on all task dimensions, but in all cases the bottom 3 points were used by less than 10 percent of those who rated the task dimension. An effective 4 point scale rather than 7 point scale appears to have operated. Table 2 indicates that the lowest scale point was used by supervisors for only 37 of the 51 dimensions, and the second lowest point was used for 48 of the 51 dimensions. The next two rating points were used on all tasks, but the use of the four lowest points was restricted to less than 10 percent of those who rated the dimension. An effective 3 point scale was operating for supervisors, but for 31 of the 51 dimensions over half the raters used the top scale value. Table 3 indicates that the second lowest scale point was used by peers for 37 of the 51 dimensions and the third lowest point was used for 50 of the 51 dimensions. The lowest and fourth lowest point was used for all dimensions but again the use of the four lowest scale values was restricted to less than 10 percent of the raters. Again, a 3 point scale seemed to be operating but for 39 of the 51 dimensions over half of the raters used the top scale value.

INTER-RATER CORRELATIONS

The relationships between incumbent, supervisor, and peer ratings are shown in Table 5. The correlations between incumbent and peer ratings on task dimensions range from .00 to .21 with over 9/10 of them .20 or below. The correlations between incumbent and supervisor ratings on task dimensions range from .00 to .36 with about 3/4 of them .20 or below. The correlation between peer and supervisor ratings on task dimensions range from .01 to .44 with slightly over half of them .20 or below. While there were a few negative calculated coefficients, none of these were statistically significant. The correlations between peer and supervisor ratings of the six general ratings ranged between .34 and .40.

The moderate correlation for the general ratings was in line with expectations but the relationships on the task dimensions was somewhat lower than expected. The implications of these findings are not yet clear.

The intercorrelations of peer and supervisors on the general ratings are shown in Table 6. As expected, the general ratings are highly intercorrelated. For supervisors the coefficients range from .76 to .87 and for peers from .73 to .86. In both cases, the patterns of intercorrelations are highly similar and in both cases the product-oriented general ratings are correlated somewhat higher with the overall performance rating than are the motivationally mediated general ratings.

CORRELATIONS BETWEEN TASK AND GENERAL RATINGS

Table 7 shows the correlations between task dimension and general ratings for the supervisors. The coefficients range from .34 to .67. Table 8 shows equivalent data for peers. These coefficients range from .23 to .67. Although there are exceptions, the coefficients for the peers tend to be somewhat lower than for the supervisors. Tables 9 and 10 show the correlations between incumbent task dimension data and supervisor and peer general ratings. These range from .00 to .21 for supervisors ratings and from .00 to .32 for peers. A few of the calculated coefficients are negative, but none of these are statistically significant. No attempts have yet been made to use multiple correlations of task data to predict the general ratings.

SUMMARY

While analyses of the data collected have not yet been completed, some implications of the use of a task level approach are indicated. In order to ensure adequate numbers at the individual task level, large original samples are required to overcome attrition of data. The task level approach does not overcome the overriding problem of restricted variance and positive skew. Inter-rater correlations on task dimensions tend to be relatively low, usually no higher than .20. Inter-rater correlations on more general ratings are higher and intercorrelations between general ratings is high. Finally, zero order correlations between task dimension and general performance ratings tend to be moderate.

Table 1

Frequency of Use of Rating Points Task Data From Incumbents

Percent of Sample*	Number of Task Dimensions (Total = 51)							
	Not Rated	Rated Using Points						
		1	2	3	4	5	6	7
0 - 10		51	51	51	30	4		
11 - 20					21	39	8	
21 - 30						8	41	5
31 - 40	12						2	16
41 - 50	17							25
51 - 60	11							3
61 - 70	8							2
71 - 80	3							
Total	51	51	51	51	51	51	51	51

*For tasks not rated, percent calculated on basis of total sample of 1034
For tasks rated, percent calculated on basis of variable samples of persons rating that task

Table 2

Frequency of Use of Rating Points Task Data From Supervisors

Percent of Sample*	Number of Task Dimensions (Total = 51)							
	Not Rated	Rated Using Points						
		1	2	3	4	5	6	7
0 - 10		37	48	51	51	13		
11 - 20						38	7	
21 - 30							42	
31 - 40	6						2	6
41 - 50	14							14
51 - 60	12							15
61 - 70	7							14
71 - 80	7							2
81 - 90	5							
Total	51	37	48	51	51	51	51	51

*For tasks not rated, percent calculated on basis of total sample of 1034
For tasks rated, percent calculated on basis of variable samples of persons rating that task

Table 3

Frequency of Use Rating Points Task Data From Peers

Percent of Sample*	Number of Task Dimensions (Total = 51)							
	Not Rated	Rated Using Points						
		1	2	3	4	5	6	7
0 - 10		51	37	50	51	24		
11 - 20						27	10	
21 - 30							39	
31 - 40							2	1
41 - 50								11
51 - 60	6							25
61 - 70	19							12
71 - 80	16							2
81 - 90	9							
91 - 100	1							
Total	51	51	37	50	51	51	51	51

*For tasks not rated, percent calculated on basis of total sample of 1034
For tasks rated, percent calculated on basis of variable samples of persons rating that task

Table 4

Distribution of Performance Ratings

Task Dimension	Incumbent Ratings			Peer Ratings			Supervisor Ratings		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
1A	662	6.14	1.20	441	6.28	1.03	624	6.40	0.92
2A	497	5.36	1.35	271	5.90	1.28	374	5.73	1.30
2B	479	5.78	1.26	280	5.96	1.33	363	5.85	1.37
3A	597	5.14	1.64	398	5.72	1.56	663	5.65	1.60
4A	398	4.97	1.68	212	5.89	1.44	266	5.74	1.29
4B	373	5.29	1.54	211	6.02	1.23	274	5.93	1.20
5A	554	5.71	1.29	380	6.02	1.22	583	6.08	1.07
6A	616	6.05	1.33	343	6.32	1.14	538	6.41	0.95
7A	650	6.11	1.12	412	6.28	1.05	603	6.31	0.99
8A	614	5.74	1.37	383	5.98	1.27	544	5.93	1.22
8B	596	5.54	1.37	379	5.94	1.22	531	5.87	1.27
8C	584	5.68	1.30	370	6.01	1.21	516	6.03	1.16
9A	696	6.12	1.00	452	6.39	0.96	664	6.34	0.97
10A	592	6.00	1.25	355	6.41	0.99	549	6.50	0.86
11A	667	5.90	1.22	435	6.08	1.27	624	6.01	1.22
11B	637	6.14	1.08	410	6.18	1.15	588	6.27	1.04
12A	561	5.82	1.38	269	6.21	1.23	419	6.23	1.06
13A	673	5.91	1.20	437	5.96	1.38	627	6.01	1.37
14A	623	6.01	1.10	381	6.22	1.14	614	6.18	1.10
15A	646	5.96	1.23	382	6.36	1.00	598	6.24	1.03
16A	606	6.08	1.19	356	6.35	0.95	529	6.41	0.92
17A	553	5.93	1.30	326	6.30	1.10	511	6.30	1.05
18A	516	5.89	1.36	298	6.50	0.96	424	6.44	1.08
19A	512	5.75	1.33	252	6.37	1.00	406	6.38	1.01
20A	502	5.95	1.42	304	6.39	1.07	503	6.40	0.94
21A	673	6.09	1.25	400	6.30	1.20	591	6.17	1.30
22A	654	5.80	1.26	371	6.24	1.10	604	6.30	0.96
23A	519	5.91	1.21	290	6.20	1.06	450	6.20	1.08
23B	462	5.82	1.29	276	6.21	1.04	430	6.15	1.07
24A	507	6.02	1.18	302	6.46	0.88	463	6.55	0.79
25A	378	6.00	1.27	216	6.41	1.05	208	6.48	0.85
26A	360	5.90	1.32	199	6.20	1.13	210	6.35	1.07
27A	349	5.93	1.46	171	6.15	1.19	215	6.36	0.91
28A	324	5.79	1.60	160	6.14	1.23	208	6.33	0.94
29A	246	5.47	1.74	90	6.08	1.23	114	6.24	1.04
30A	272	5.40	1.80	110	6.21	1.10	139	6.14	1.07
31A	314	5.78	1.60	139	6.35	1.12	190	6.37	0.97
32A	293	5.69	1.61	134	6.19	1.16	178	6.38	1.00
33A	322	5.52	1.66	134	6.08	1.31	185	6.23	1.11
34A	474	5.90	1.42	216	6.19	1.18	361	6.06	1.19
35A	458	5.79	1.46	243	6.19	1.22	405	6.15	1.13
36A	489	5.78	1.43	224	6.14	1.23	356	5.91	1.28
37A	436	5.61	1.50	180	6.18	1.23	306	6.02	1.14
38A	435	5.88	1.44	189	6.34	1.11	323	6.33	0.99
39A	486	5.79	1.48	226	6.15	1.27	357	5.93	1.30
40A	524	6.35	1.11	269	6.60	0.99	443	6.66	0.74
41A	548	5.94	1.43	306	6.34	1.21	468	6.37	1.06
42A	447	5.81	1.30	160	6.36	1.03	288	6.43	0.87
43A	580	5.99	1.28	349	6.34	1.00	537	6.27	1.07
43B	511	5.86	1.14	319	6.25	1.08	506	6.18	1.05
44A	567	6.32	1.14	321	6.55	1.04	541	6.62	0.87

Table 4 (cont'd)

General Ratings	Incumbent Ratings			Peer Ratings			Supervisor Ratings		
	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
*G1				551	79.36	16.53	799	80.86	15.03
G2				632	79.78	17.54	850	81.84	15.87
G3				632	81.31	17.48	849	83.36	15.45
G4				630	78.10	20.84	849	80.65	19.53
G5				630	80.81	19.64	851	81.68	19.46
G6				630	81.98	18.73	852	83.02	17.51

*G1 = Overall Performance
G2 = Amount of Work Produced
G3 = Quality of Work Produced
G4 = Willingness To Do More Than His Share of the Work
G5 = Willingness To Start and Continue Work on His Own Without The Need for Prodding
G6 = Willingness To Share His Skill and Knowledge With Peers and Subordinates

Table 5

Relationships Between Incumbent, Peer, and Supervisors Reported Task Performance

Task Dimension No.	Incumbent vs Peer		Incumbent vs Supervisor		Peer vs Supervisor	
	r	n	r	n	r	n
1A	06	357	15	499	26	366
2A	18	165	21	236	26	140
2B	-01	159	16	214	39	139
3A	09	275	11	435	12	304
4A	06	108	22	139	39	85
4B	00	102	10	129	44	86
5A	21	236	12	355	31	244
6A	13	275	12	424	14	266
7A	17	343	13	497	29	343
8A	16	319	24	447	28	312
8B	20	303	26	424	35	300
8C	15	292	20	407	39	288
9A	07	380	05	528	24	369
10A	16	282	11	420	08	278
11A	12	373	21	513	35	373
11B	04	339	12	466	30	333
12A	06	196	28	303	12	162
13A	08	366	13	519	23	367
14A	12	294	05	455	18	309
15A	06	318	05	490	12	321
16A	14	286	15	419	10	282
17A	10	243	11	362	18	238
18A	02	235	10	311	18	218
19A	04	176	06	283	12	169
20A	08	213	03	321	-02	224
21A	-01	334	13	475	22	319
22A	05	296	07	441	08	282
23A	13	210	11	321	22	209
23B	14	175	09	270	28	190
24A	17	212	02	303	12	207
25A	10	122	20	106	-02	79
26A	-02	108	22	98	13	66
27A	05	136	10	177	06	118
28A	01	121	16	157	13	111
29A	08	51	-02	74	-06	38
30A	02	66	12	81	16	51
31A	15	110	00	149	03	96
32A	-06	92	05	133	03	84
33A	03	103	18	146	15	82
34A	12	170	13	279	28	162
35A	19	168	22	261	41	169
36A	20	179	26	279	32	164
37A	10	131	29	226	32	112
38A	11	134	18	234	26	126
39A	19	183	36	282	38	166

Table 5 (cont'd)

Task Dimension No.	Incumbent vs Peer		Incumbent vs Supervisor		Peer vs Supervisor	
	r	n	r	n	r	n
40A	10	188	07	303	-01	173
41A	09	237	05	354	12	224
42A	21	90	08	162	-05	61
43A	13	249	07	389	10	241
43B	09	200	09	330	17	210
44A	08	218	27	361	04	223
*G1					34	455
G2					40	548
G3					37	548
G4					34	545
G5					39	547
G6					35	547

Summary of Intercorrelations

Range of r	No. of Task Dimensions I vs P	No. of Task Dimensions I vs S	No. of Task Dimensions P vs S
00-05	12	9	7
06-10	17	10	6
11-15	11	14	11
16-20	9	6	4
21-25	2	6	4
26-30		5	8
31-35			4
36-40		1	5
41-44			2

*G1 = Overall Performance
G2 = Amount of Work Produced
G3 = Quality of Work Produced
G4 = Willingness To Do More Than His Share of the Work
G5 = Willingness To Start and Continue Work on His Own Without The Need for Prodding
G6 = Willingness To Share His Skill and Knowledge With Peers and Subordinates

Table 6

Relationships Between Peer and Supervisor Reported Performance on General Factors

	Peer Ratings				
	G2	G3	G4	G5	G6
G1	549 84	549 86	549 79	549 81	548 76
G2		630 81	629 77	628 79	628 73
G3			628 76	628 77	628 75
G4				629 87	626 78
G5					627 81

	Supervisor Ratings				
	G2	G3	G4	G5	G6
G1	798 87	798 87	795 82	797 83	798 79
G2		848 85	845 82	848 84	849 78
G3			845 78	847 80	848 76
G4				848 91	847 80
G5					849 83

Peer General Ratings	Supervisor General Ratings					
	G1	G2	G3	G4	G5	G6
G1	455 34	478 39	478 37	477 37	479 40	479 35
G2	521 34	548 40	548 35	547 34	549 39	549 35
G3	521 35	548 41	548 37	547 37	549 40	549 34
G4	519 29	546 34	546 31	545 34	547 36	547 31
G5	519 32	546 36	546 33	545 37	547 39	547 32
G6	519 34	546 37	546 33	545 36	547 38	547 35

G1 = Overall Performance

G2 = Amount of Work Produced

G3 = Quality of Work Produced

G4 = Willingness To Do More Than His Share of the Work

G5 = Willingness To Start and Continue Work on His Own Without the Need for Prodding

G6 = Willingness To Share His Skill and Knowledge with Peers and Subordinates

NOTE: First entry in each cell = number of cases. Second, underlined entry in each cell = correlation coefficient without decimal point

Table 7

Relationships Between Supervisor Task Performance Ratings and General Performance Ratings

Task Dimension Number	G1		G2		G3		G4		G5		G6	
	r	n	r	n	r	n	r	n	r	n	r	n
1A	58	561	53	599	57	599	50	595	53	597	53	598
2A	58	338	54	358	52	358	56	356	56	358	64	357
B	65	330	66	348	63	348	67	346	67	348	65	347
3A	46	601	45	637	44	637	42	633	43	635	44	636
4A	63	243	64	251	63	251	60	250	62	251	67	250
B	63	248	66	258	66	258	60	257	62	258	63	257
5A	62	521	60	551	58	552	56	549	60	551	61	551
6A	53	486	51	518	52	518	47	514	51	516	50	517
7A	58	548	51	579	60	579	47	576	51	578	47	578
8A	60	490	58	521	60	521	51	518	56	520	54	520
B	60	479	58	509	59	509	50	506	57	508	54	508
C	62	466	62	495	59	495	54	492	58	494	55	494
9A	57	607	59	639	59	639	54	635	56	637	57	638
10A	50	493	46	524	47	524	42	520	47	522	51	523
11A	56	563	57	595	59	595	53	591	55	593	49	594
B	59	530	60	562	59	562	59	559	61	561	55	561
12A	51	382	52	402	52	402	43	398	47	400	49	401
13A	46	565	45	598	50	598	45	595	46	597	43	597
14A	56	558	53	586	59	586	50	582	52	584	55	585
15A	55	544	55	573	55	573	52	569	54	571	54	572
16A	47	479	46	503	48	503	44	500	45	501	46	502
17A	53	471	53	494	51	494	52	492	55	493	49	493
18A	44	388	47	413	46	413	46	410	45	412	46	412
19A	60	371	59	392	59	392	55	390	57	391	57	391
20A	40	450	35	479	36	479	34	476	40	478	42	478
21A	42	536	44	567	39	567	39	564	43	566	43	566
22A	49	541	43	574	45	574	39	571	45	573	50	573
23A	55	402	53	428	57	428	55	425	58	427	54	427
B	60	384	59	408	58	408	60	405	63	407	57	407
24A	41	418	42	446	41	446	37	444	40	445	42	445
25A	49	190	48	203	53	203	43	203	42	203	47	203
26A	60	194	50	206	59	206	47	206	50	206	56	206
27A	70	188	67	203	69	203	66	204	68	204	64	204
28A	65	182	63	197	65	197	60	198	62	198	61	198
29A	54	101	59	111	66	111	47	112	52	112	64	112
30A	56	120	55	131	61	132	55	132	61	132	55	132
31A	64	163	64	178	71	178	60	179	62	179	66	179
32A	59	155	57	169	65	169	54	170	56	170	63	170
33A	53	163	52	175	57	175	53	176	55	176	52	176
34A	54	329	50	343	49	343	41	341	43	343	50	342
35A	48	369	42	391	44	391	37	389	38	391	44	390
36A	55	323	52	337	51	337	44	335	47	337	47	336
37A	53	276	49	288	52	288	45	286	47	288	50	287

Table 7 (cont'd)

Task Dimension Number	G1		G2		G3		G4		G5		G6	
	r	n	r	n	r	n	r	n	r	n	r	n
38A	53	293	47	305	53	305	44	303	45	305	50	304
39A	55	325	50	339	52	339	43	337	44	339	46	338
40A	34	403	30	425	36	425	34	423	35	424	41	424
41A	42	428	41	453	41	453	44	449	45	451	44	452
42A	50	261	49	278	51	278	46	276	44	277	44	277
43A	55	490	55	514	54	514	54	510	56	512	55	513
B	56	464	56	487	56	487	55	483	56	485	54	486
44A	31	493	31	516	31	516	30	512	34	514	35	515

G1 - Task versus Overall Rating
G2 - Task versus Amount of Work Produced
G3 - Task versus Quality of Work Produced
G4 - Task versus Willingness to Do More Than His Share of Work
G5 - Task versus Willingness to Start and Continue Work on His Own Without Need for Prodding
G6 - Task versus Willingness to Share His Skill and Knowledge With Peers and Subordinates

Table 8

Relationships Between Peer Task Performance Ratings and General Performance Ratings

Task Dimension Number	G1		G2		G3		G4		G5		G6	
	r	n	r	n	r	n	r	n	r	n	r	n
1A	49	365	45	413	47	412	45	411	46	412	41	414
2A	61	228	48	257	51	257	55	255	55	256	56	258
B	67	238	55	265	56	264	62	263	66	264	55	266
3A	37	335	28	378	34	376	35	376	33	376	26	377
4A	54	174	47	197	51	197	52	195	55	196	50	198
B	59	173	46	198	50	197	56	195	59	195	44	198
5A	53	314	50	356	45	356	47	353	52	354	44	356
6A	35	286	32	321	33	320	27	318	29	319	24	322
7A	54	340	50	386	49	385	45	383	48	384	42	387
8A	52	320	48	361	54	361	44	358	47	359	47	362
B	53	317	49	358	52	357	41	355	48	356	43	359
C	59	311	52	350	57	349	51	348	53	349	53	351
9A	47	376	43	427	45	427	38	424	40	425	40	427
10A	28	295	28	332	28	331	23	329	28	330	21	333
11A	49	357	47	408	49	407	41	405	40	406	43	409
B	52	337	48	384	51	383	46	381	50	382	46	385
12A	42	216	43	257	44	258	32	254	35	255	30	258
13A	31	359	28	410	33	410	31	407	30	408	31	411
14A	44	318	44	363	48	362	39	360	41	361	43	363
15A	42	324	38	363	43	362	37	360	43	361	42	364
16A	37	298	37	330	39	331	30	328	32	329	33	331
17A	42	270	39	308	42	306	37	306	41	306	42	307
18A	30	245	30	281	32	280	33	279	34	279	31	280
19A	46	213	39	244	44	245	37	242	44	243	39	244
20A	42	243	39	283	38	283	37	282	43	282	34	282
21A	40	330	36	381	36	380	38	379	37	379	34	380
22A	43	302	42	352	40	353	37	349	42	350	37	352
23A	51	243	48	274	41	274	42	271	50	272	42	274
B	55	231	48	260	44	260	45	258	54	259	42	260

G1 = Task versus Overall Rating
G2 = Task versus Amount of Work Produced
G3 = Task versus Quality of Work Produced
G4 = Task versus Willingness to Do More Than His Share of Work
G5 = Task versus Willingness to Start and Continue Work on His Own without Need for Prodding
G6 = Task versus Willingness to Share His Skill and Knowledge with Peers and Subordinates

Table 8 (Continued)

Task Dimension Number	G1		G2		G3		G4		G5		G6	
	r	n	r	n	r	n	r	n	r	n	r	n
24A	38	248	31	288	37	286	27	286	28	286	27	288
25A	34	173	27	206	24	206	26	204	28	204	25	205
26A	36	160	38	191	32	190	26	188	32	188	31	190
27A	39	132	35	153	36	155	35	153	36	154	29	155
28A	39	122	36	144	37	145	34	143	37	143	29	145
29A	21	73	21	81	18	81	17	80	18	80	10	81
30A	34	81	27	98	22	98	19	97	29	97	18	98
31A	36	104	25	122	22	122	20	121	26	121	02	122
32A	30	100	19	118	24	118	09	117	11	117	-01	118
33A	39	105	38	122	41	122	33	121	30	121	11	122
34A	37	176	34	203	35	202	18	201	29	202	24	204
35A	38	199	40	232	35	231	26	229	30	230	30	233
36A	43	177	48	210	41	209	33	207	39	208	35	211
37A	48	139	48	169	47	167	31	166	39	166	32	169
38A	31	149	36	176	28	174	19	173	23	173	22	176
39A	39	182	43	212	42	211	29	209	34	210	33	213
40A	29	219	29	253	31	253	18	250	23	250	20	253
41A	31	251	22	288	36	289	34	287	28	288	30	289
42A	32	133	30	155	27	155	21	154	24	154	28	155
43A	39	290	35	334	39	333	34	333	29	333	34	334
B	42	266	35	306	41	305	34	304	32	304	33	306
44A	23	265	19	306	23	306	23	304	23	304	17	306

Table 9

RELATIONSHIPS BETWEEN INCUMBENT TASK PERFORMANCE RATINGS AND PEER GENERAL PERFORMANCE RATINGS

Task Dimension Number	G1		G2		G3		G4		G5		G6	
	r	n	r	n	r	n	r	n	r	n	r	n
1A	11	369	13	418	07	418	02	415	05	416	05	419
2A	11	277	12	317	12	317	01	315	05	316	13	318
B	07	269	13	307	13	307	05	305	07	306	11	308
3A	17	341	14	389	14	389	10	387	15	388	16	389
4A	11	224	13	252	14	252	09	250	08	251	13	253
B	15	207	11	233	14	233	05	231	09	232	17	234
5A	16	308	16	352	16	352	06	349	10	350	19	352
6A	12	339	12	388	09	388	04	385	04	386	10	389
7A	20	365	15	412	17	412	09	409	13	410	17	413
8A	19	342	17	391	18	391	10	388	13	389	19	391
B	22	331	22	378	20	378	11	375	16	376	21	378
C	20	328	20	372	19	372	11	369	16	370	21	372
9A	16	396	12	449	09	449	05	446	09	447	10	450
10A	11	334	10	377	09	377	05	374	05	375	09	378
11A	13	377	10	426	09	426	07	423	08	424	11	427
B	10	359	08	405	03	405	04	402	08	403	09	406
12A	13	315	19	359	12	359	13	357	12	358	17	360
13A	12	381	09	433	06	433	04	430	05	431	02	433
14A	17	346	15	395	15	395	14	393	13	394	18	395
15A	19	362	14	413	13	413	07	410	08	411	16	414
16A	14	344	09	386	09	386	08	383	03	384	10	387
17A	15	314	11	355	11	355	11	353	10	354	14	355
18A	11	308	07	350	08	349	05	348	10	348	12	349
19A	15	292	14	333	12	332	10	331	13	331	19	332
20A	08	276	08	315	05	315	01	313	00	314	07	316
21A	10	380	08	431	09	431	07	429	04	430	11	432
22A	12	362	11	414	09	414	05	412	11	413	08	414
23A	03	291	05	330	01	330	02	328	02	329	-02	330
B	09	267	14	297	08	297	06	295	10	296	06	297

G1 = Task versus Overall Rating
G2 = Task versus Amount of Work Produced
G3 = Task versus Quality of Work Produced
G4 = Task versus Willingness to Do More Than His Share of Work
G5 = Task versus Willingness to Start and Continue Work on His Own without Need for Prodding
G6 = Task versus Willingness to Share His Skill and Knowledge with Peers and Subordinates

Table 9 (Continued)

Task Dimension Number	G1		G2		G3		G4		G5		G6	
	r	n	r	n	r	n	r	n	r	n	r	n
24A	04	284	06	323	04	323	02	321	02	322	07	324
25A	04	208	07	245	08	245	-03	243	00	244	08	245
26A	18	195	19	230	20	230	09	228	13	229	18	230
27A	-06	184	01	215	01	214	-08	213	-03	213	04	214
28A	01	171	05	201	06	200	-06	199	00	199	06	200
29A	02	128	08	152	07	151	-04	150	06	150	02	151
30A	04	139	08	167	12	166	-04	165	03	165	08	166
31A	00	164	02	193	04	193	-04	192	01	192	04	193
32A	-03	153	05	179	02	179	-04	178	00	178	10	179
33A	13	168	07	196	16	196	07	195	07	195	16	196
34A	13	268	20	300	18	300	10	298	13	299	19	301
35A	25	269	32	300	30	300	24	298	29	299	26	301
36A	13	272	18	309	18	309	08	307	13	308	15	310
37A	15	245	21	279	20	279	11	277	16	278	19	280
38A	15	238	17	270	20	270	08	268	12	269	16	271
39A	14	269	18	306	17	306	08	304	13	305	16	307
40A	08	292	04	333	05	333	04	331	05	332	04	333
41A	14	299	08	341	13	341	13	339	10	340	12	341
42A	11	242	14	275	13	275	05	273	10	274	18	276
43A	14	327	17	367	14	367	10	365	11	366	15	367
B	06	284	12	317	09	317	04	315	05	316	06	317
44A	18	313	16	359	13	359	08	357	11	358	13	360

Table 10

Relationships Between Incumbent Task Performance Ratings and Supervisors General Performance Ratings

Task Dimension Number	G1		G2		G3		G4		G5		G6	
	r	n	r	n	r	n	r	n	r	n	r	n
1A	17	519	15	547	17	548	08	547	14	548	12	549
2A	17	398	21	417	17	418	17	417	18	418	16	420
B	15	382	15	401	14	402	10	401	09	402	11	404
3A	12	467	13	495	13	496	13	496	16	497	13	497
4A	11	323	09	335	11	336	04	336	09	337	08	338
B	11	300	08	312	08	313	00	313	04	314	08	315
5A	13	442	14	464	13	465	08	464	11	465	12	466
6A	16	489	19	513	19	514	14	513	15	514	14	515
7A	09	518	12	543	12	544	06	543	11	544	09	545
8A	11	484	10	509	12	510	06	509	11	510	09	511
B	12	472	13	494	15	495	09	494	11	495	12	496
C	18	464	17	486	19	487	12	486	15	487	16	488
9A	12	556	14	584	15	585	10	584	14	585	12	586
10A	12	469	12	490	14	491	09	490	12	491	10	492
11A	10	527	12	552	11	553	07	552	11	553	10	554
B	03	507	05	530	07	531	03	530	07	531	02	532
12A	10	451	11	473	11	474	05	473	07	474	06	475
13A	11	533	12	558	16	559	10	558	11	559	12	560
14A	14	495	16	520	17	521	11	520	13	521	14	522
15A	08	516	09	539	10	540	07	539	09	540	10	541
16A	10	482	10	505	12	506	08	505	11	506	09	507
17A	11	432	13	458	13	458	12	456	14	457	10	459
18A	05	400	09	421	06	422	03	421	04	422	04	423
19A	05	398	06	418	09	419	02	418	06	419	05	420
20A	-01	397	01	419	02	420	-01	418	-04	419	-01	421
21A	09	534	10	561	11	562	11	560	10	561	07	563
22A	09	516	10	543	13	544	06	542	08	543	08	545
23A	16	417	16	438	17	439	13	437	14	438	13	439
B	13	375	17	391	16	392	12	391	11	392	12	393
24A	01	403	01	421	04	422	-01	420	02	421	00	422
25A	08	294	05	303	10	304	03	304	05	304	03	305
26A	16	281	12	292	20	293	10	293	14	293	10	294
27A	07	269	07	287	12	287	02	290	04	289	05	290
28A	09	247	09	265	16	265	03	268	05	267	09	268
29A	02	189	00	203	07	203	-04	205	-04	204	-01	205
30A	05	208	05	225	11	224	03	227	-02	226	02	227
31A	05	239	02	256	11	256	02	259	00	258	03	259
32A	05	227	08	242	12	242	02	245	04	244	09	245
33A	11	246	10	264	17	264	08	267	08	266	08	267
34A	09	377	09	389	13	390	08	388	10	389	07	390

Table 10 (cont'd)

Task Dimension Number	G1		G2		G3		G4		G5		G6	
	r	n	r	n	r	n	r	n	r	n	r	n
35A	13	366	14	378	14	379	12	378	16	379	10	380
36A	10	390	09	403	12	404	10	403	11	404	07	405
37A	16	351	16	362	18	363	14	361	17	362	14	363
38A	16	345	11	356	17	357	13	355	15	356	10	357
39A	14	387	13	400	15	401	11	399	11	400	09	401
40A	07	409	09	430	12	431	06	430	08	431	05	432
41A	03	425	04	445	06	446	03	445	04	446	00	447
42A	07	351	07	372	08	373	03	371	05	372	03	373
43A	13	458	13	479	12	480	15	479	18	480	18	481
B	03	407	03	425	04	426	01	424	04	425	03	427
44A	07	455	07	478	10	479	05	478	06	479	03	480

G1 - Task versus Overall Rating
G2 - Task versus Amount of Work Produced
G3 - Task versus Quality of Work Produced
G4 - Task versus Willingness to Do More Than His Share of Work
G5 - Task versus Willingness to Start and Continue Work on His Own Without Need for Prodding
G6 - Task versus Willingness to Share His Skill and Knowledge With Peers and Subordinates

SCALING TECHNIQUES TO REDUCE RATER LENIENCY ON NAVY ENLISTED PERFORMANCE EVALUATION FORMS[1]

Marjorie H. Royle and David W. Robertson
Navy Personnel Research and Development Center
San Diego, California 92152

The literature on performance evaluation abounds with frequent mention of terms to which sources of evaluation error are attributed. Some of the most commonly employed terms are leniency, central tendency, halo, and skewness. In the discussions in many of the studies, there appears to be an assumption that these terms represent actual phenomena which have been empirically established. Far too few of the papers raise serious questions regarding the justification for the assumption (implicit or explicit) of the actual existence of these phenomena, or the criteria for establishing their existence.

The first of the error terms listed above, leniency, is usually defined as the tendency of the rater to evaluate the ratee more favorably than the actual level of the ratee's performance. Sharon and Bartlett (1969) cite a number of examples from the literature which suggest that there is some confusion regarding operational definitions of this term.

[It might be mentioned, parenthetically, that there already is some degree of confusion inherent in performance evaluation studies on Navy subjects, in the light of the Navy term for an occupational specialty, Rating, and the Navy term for a specific Rating at a specific pay grade, Rate. For example, the Navy Rating, Aviation Machinist Mate, at the Pay Grade 5 level, is the Rate of Aviation Machinist Mate Second Class. Given the existence of such homonymously equivocal terms, the term "evaluation" is usually employed (in Navy studies), rather than "Rating," and the terms, rater and ratee, are used only with some reluctance.]

Some of the methods cited by Sharon and Bartlett (1969), as inappropriate for measuring leniency, include the following two: First, the proximity of the mean of the evaluations to the mid-point of the scale employed--inadequate because it assumes that the true mean should be at the mid-point. Second, the skewness of the distribution, is inadequate, because the true distribution may be skewed, particularly if the ratees are supervisory personnel. A negative skew may simply reflect a situation in which all but the superior performers were eliminated in previous selections (Bass, 1956).

---

[1] Paper presented at the 16th Annual Conference of the Military Testing Association, Oklahoma City, Oklahoma, 21-25 October, 1974.

Studies comparing the graphic type scale with other types of scales (e.g., forced-choice and adjective check-list) typically find that the graphic type scale is least effective in resisting leniency or excessively favorable evaluations. In studies comparing evaluations under different conditions of the uses of the evaluation data, raters typically are found to mark more favorably under conditions in which the marks are to be used in personnel actions affecting the ratee, or in which the marks must be justified to the ratee, than under other conditions (e.g., for research purposes only, to validate training curricula, or with no feedback to the ratee). The Sharon and Bartlett (1969) study, in a two by eight design, compared graphic and forced-choice scales across four conditions of use of marks. They made no assumptions regarding leniency, but rather, compared only the shift in favorability by scale type or condition. For example, they found that the graphic scale was less effective than the forced-choice scale in upward shifts in favorability under conditions of the marks affecting the ratee, or in the rater having to justify the marks to the ratee. However, it was also found that there was greater resistance (by the raters) to the forced-choice than to the graphic scale, in terms of raters leaving items blank.

In another study, involving shift in favorability, Gordon (1970) found that the accuracy of evaluations is not related to the favorability of the rater's impression (i.e., the purported halo effect) of the ratee. This suggests that the ability to evaluate accurately is not impaired by the tendency to view the ratee in a generally favorable manner.

The conclusions of most studies would seem to discourage the use of general graphic scales for supervisory evaluations to acquire data to be used in personnel actions affecting the ratee. The two purposes of this paper are:

1. To suggest some justifications for the use of graphic scales for supervisory evaluations, and

2. To report the results of some efforts to achieve desired distributions of marks.

The size of the list of disadvantages of employing graphic scales and supervisory ratings, appears to exceed the size of the list of advantages. Some of the weaknesses typically suggested, mostly in terms of reasons why raters submit excessively high marks, are:

1. A natural human reluctance to evaluate any one as "inferior."

2. The rater's desire and need for continued hard work from the ratee (i.e., motivating ratee with high marks).

3. A reluctance to "face" the ratee with low marks.

4. A suspicion (frequently confirmed) that other similar units are also marking excessively high, so that this rater must mark high enough to protect his ratees.

5. The press of periodic administrative inspections when each unit is scrutinized for its proportion of personnel advanced and for morale in general, since these considerations reflect in turn on the evaluation of the top leaders of the unit, and

6. In the special case of supervisory personnel, a tradition (at least in the Navy) that any evaluation mark short of "walks on water" is inferior.

Since important personnel selection decisions directly affecting the career of the individual must be based on some measure of job performance, and since a measure of production output is frequently neither available, relevant nor adequate, particularly in assessing military or supervisory performance, the requirement for judgmental-type evaluations appears to be unavoidable. Alternative approaches to graphic scales have been demonstrated to be superior, but they too have their disadvantages. Three types of alternatives, and their disadvantages, are:

1. Forced-choice tetrads and adjective check-lists, which are more resistant to leniency than graphic scales, but are disliked and resisted by raters, because the raters are not aware of the weighted values of each item and do not want to submit marks affecting the career of their ratees without an awareness of the values of the items marked.

2. Objectively anchored evaluation scales (e.g., behavioral anchors--Smith and Kendall, 1963) which provide explicit criteria for narrowly defined types of job behavior within specific occupational specialties, and as such, are less susceptible (than scales with more general descriptors) to error variance in marking standards across units or departments. However, the procedures employed to develop the anchors are quite extensive and would be enormously expensive to develop and update scales for each of the hundreds of occupational specialties in large organizations (such as the military services). In a feasibility study to develop general anchors which would apply to all Navy occupational specialties (i.e., Ratings) at the Pay Grades 5 and 6 levels (Robertson, 1972), it was found that:

a. The sets of descriptors were incomplete (i.e., numerous scale points and pairs of adjacent scale points had no descriptors).

b. Some of the descriptors did not appear to apply to the same continuum or dimension, and

c. The highest level of performance on each scale (i.e., the top descriptor) appeared to describe a level of performance even lower than the modal response of the operational scale.

3. Evaluation of job samples of behavior in a performance assessment center where conditions are controlled and evaluations can be conducted jointly by staff experts and visiting managers. The disadvantages of this approach include: The expense of operating the center, the conflict in time consumed and in availability of taking manager-ratees off the job, and the tenuousness of the assumption that the behavior specified in the assessment center adequately taps the real-world behavior of the job. In a study which compared evaluations in an assessment center with follow-up evaluations on the job (Thomson, 1970), it was found that the supervisory evaluations on the job were more lenient and of poorer quality in terms of restricted variance. One might seriously question the meaningfulness of such a comparison of behaviors in two very different work environments.

Thus, in spite of its many weaknesses there appears to be a need to employ, for the foreseeable future, the relatively inexpensive, quickly administered, graphic scale with general descriptors, in acquiring supervisory evaluations. Having embarked on this course, two of the objectives of performance evaluation research in the Navy, in improving the effectiveness of general graphic scales, are to:

1. Direct the mean away from the high end of the scale, toward the middle of the marking scale, and

2. Maximize variance.

No assumptions were made regarding leniency (i.e., in terms of distributions differing from some specified criterion). Rather, various scaling characteristics were tested to determine which were most effective in resisting pile-up of marks toward the high end of the scale.

A three by four design was tested on types of conversion codes for the marks and types of scaling characteristics respectively (Robertson et al., 1972a). The ratees in the Navy-wide trial administration were in seven occupational specialties (considered reasonably representative of all occupational areas, e.g., Electronics Technician, Personnelman, etc.) at the Pay Grades 5 and 6 level of seniority.

The design features which were varied on the four basic alternative trial forms were, essentially, scale descriptors and scale length. The design features which were held constant on the basic forms were: The modal anchor or "T-zone," scale direction, required justification for extreme high or low marks, and evaluation items and item definitions. The four basic scale designs are displayed in Figure 1. (A large number of designs were developed. The four basic forms selected for trial administration, Forms 1a, 2a, 3 and 5, were estimated to be the most likely in resisting leniency, and the most unique in design characteristics among each other.)

All of the forms display essentially relative type descriptors (as distinguished from "objective" or criterion type anchors). Since personnel at the Pay Grades 5 and 6 levels have already competed successfully for advancement two or three times, up through the lower pay grades, they represent a first-line supervisor or experienced technician group, and are generally high quality performers. Therefore, the use of any terms (in the T-zone) which explicitly or even connotatively describe an "average" performer was carefully avoided. In fact, one of the very specific objectives of Navy performance evaluation research is to develop forms and procedures which permit the unit commands to acknowledge the superior quality of performance which is typical of ratees in leadership and experienced technician positions (Robertson, 1972).

Form 1a provides for eight columns of various levels of *effectiveness*, and two columns of unsatisfactory performance. If the tendency for evaluators to mark at or near the top of the scale reflects the high quality of the ratee, rather than the "leniency" of the rater, construction of scale points which emphasize that high quality should be resistant to marks piling up at the top of the scale. Therefore, *all* of the verbal descriptors on the forms are, in general, high quality terms (for these high quality ratees). For example, the top three columns of Form 1a, provide descriptors of "extremely effective *and* excels over": (1) "virtually all," (2) "nearly all," and (3) "most others of the same Rate."

Forms 2a and 3 display "variable-cumulative" scales (e.g., top 5%, top 15%, etc.) with Form 2a dichotomized in top and bottom cumulative percentages, and Form 3, trichotomized. Form 5 is physically the most similar to the then existing operational form (NAVPERS 792) (i.e., of five double columns), although the column descriptors were completely modified and a T-zone was added. The purpose of the double columns in Form 5 (and the trichotomized areas of Form 3) was to provide the rater with the opportunity for an initial broad choice, and then a finer upper or lower differentiation.

The use of a T-zone display appeared to be highly effective in the development of the evaluation form for Chief Petty Officers, Pay Grades 7-9 (Robertson, 1969). However, the finding that the T-zone was resistant to rater leniency was confounded, because the experimental form displaying the T-zone also required justification comments in a narrative section (for extremely high or low marks), whereas the alternative experimental form without the T-zone required no such justification. In order to confirm the effectiveness of the T-zone in the Pay Grades 5-6 study, a Form 2b, which was identical in all respects to Form 2a except that it displayed no T-zone, was also trial administered.

The coding procedure employed in the data storage of evaluation marks is extremely important, because many individual personnel decisions are based on criteria expressed in terms of these codes. Some of these important selection decisions include eligibility for: Honorable Discharge, Good Conduct Medal, Reenlistment, and Advancement. Thus, it appears that knowledge of the *use* of the evaluation data, in terms of the codes employed, may influence the evaluation procedures, particularly in regard to the problem of leniency. The following three types of coding procedures were investigated (and are displayed in Table 1):

1. Numerical (old) - a discontinuous set of codes ranging from 1.0 to 4.0, which have been used for storing enlisted performance evaluation data, and have been the dimension in which various individual personnel actions are defined, for many years in the Navy.

2. Numerical (new) - a continuous set of codes, ranging from 0 to 9, which were used to perform the analysis in the previous E-7/8/9 study, the E-5/6 study, and for the storage of operational evaluation data.

3. Graphic - a set of alpha-numeric codes which reflect or connote the position of the evaluation mark on the scale (e.g., for a mark in the "top 35%" column, a code of T35).

For the trial administration of the alternative forms and coding procedures, commands were instructed to perform the evaluation as if they were actual (i.e., operational) evaluation documents, including review up the chain of command. In the Navy enlisted system, the performance evaluation form is completed initially by the ratee's immediate supervisor, and then reviewed at various intermediate levels, concluding with the Commanding Officer or his representative. Changes in marks or written supporting materials can and do occur at each step of review. In the sampling design employed, no unit command was sampled with ratees in more than one of the scale-by-code subgroups.

The results of the trial administrations were analyzed for mean differences and distributional differences, particularly for pile-up of marks in the top three columns of the scales. Multiple analyses of variance were performed in order to test for the interaction of coding by scale by occupational specialty (i.e., Navy Rating).

The results indicated that the trial forms were substantially more effective in resisting leniency than were the operational forms, and that certain scaling and coding designs were superior to others. In comparing otherwise similar scales, differing only in the presence or absence of a modal anchor (i.e., T-zone), the results reflect the superiority of the T-zone. Table 2 presents similar findings from both the Pay Grades 7-9 and 5-6 studies, with the T-zone scales yielding the lower means. (In the Pay Grades 5-6 study, the comparison was between Forms 2a and 2b.)

Many studies throughout the literature report mean differences resulting from the rater's knowledge of the use of the marks. The means of "For Research Purpose Only" marks are generally lower than those which actually affect the ratee. Thus, it is of critical interest whether the distribution of marks from the trial administration of a prospective operational scale will remain stable when the scale is operationally implemented. In the Pay Grades 7-9 and 5-6 studies, raters were instructed to evaluate the ratee as if the marks were to be used operationally.

Figure 2 presents a comparison of Pay Grade 7 data from three different analyses--of the old operational scale, the trial scale of the new operational form, and the new operational scale. The results suggest that the "as if operational" instructions were effective in yielding the same distributions which could be expected if a given trial scale were selected for the new operational design.

The results of the multiple analyses of variance, Table 3, indicate that all three main effects--coding, scaling type and job specialty--were significant. (However, the job specialty effect was of limited interest, since most personnel actions, especially advancement competition and duty assignment, are within job specialty.) Also, some extremely interesting, statistically significant, interactions between coding and scaling were identified. Figure 3 presents the means, by coding and scaling types, for Pay Grades 5 and 6 combined, for two of the evaluation items--Individual Productivity and Directing. One of the most striking findings of the study was the stultifying effect of the old operational 1.0-4.0 codes--even when applied to the new scales. Thus, this code produced a relatively flat and in-the-middle line, regardless of scaling characteristics. It has frequently been suggested that the primary reason for the progression of marks towards the high end of the marking scale over time is attributable to familiarity with the form. Nonetheless, the present finding suggests that a more valid reason may be: Familiarity with the marking codes and the uses of those codes in terms of the criteria for various personnel actions. From Figure 3 it is apparent that the graphic or the 0-9 codes, combined with the Form 2a or Form 5 scale, are most resistant to leniency. Also, supervisory effectiveness (i.e., the Directing item) appears generally to be less impressive to the rater than the ratee's individual output.

Table 4 presents the sum of the percentages of marks entered in the top three columns of each marking scale and coding type. Generally, the 1.0-4.0 codes were least resistant to rater leniency, and the graphic codes were most resistant. Again, differences among scale types were less clear-cut, but Forms 2a and 5 were generally better than Forms 1a and 3. Thus, it appears that there is little difference in effectiveness between the descriptive and cumulative percentage scales of Forms 5 and 2a respectively. This finding is in contrast to the results of the Pay Grades 7-9 study in which the cumulative percentage descriptors were the most effective in holding down the pile up of marks at the high end of the scale. The finding suggests that, at the Pay Grades 5-6 level, qualitative descriptors may still be more effective than categorical descriptors or percentages, while at the 7-9 level, an adequate number or type of qualitative superlatives does not exist, and percentage descriptors are then superior.

Form 3 was the least effective, and also had the strongest interaction effect, particularly with the 1.0-4.0 code. The form allocated only one column to the T-zone, and there were no explicit descriptors of superior quality displayed in the T-zone. The resulting marks suggest that the raters avoided the middle area of the trichotomy, being reluctant to evaluate ratees as "average," and utilized primarily the top three categories. Further, since the mean of the 1.0-4.0 coding type was the lowest (on Form 3) of the three coding types, the raters' familiarity with the old code may have been the predominant anchoring influence.

The unit commands which participated in the trial administration were provided with comment and suggestion questionnaires for all levels of review in the chain of command. Fleet reaction generally was quite favorable. Most negative criticism was directed to the Optical Character Reader (OCR) mode, since automated forms were evaluated in the same trial administration. (The feasibility of automated forms and computer-generated data displays is discussed in Robertson et al., 1972a and b.) Much less criticism was directed towards the substantive design features. However, there are always exceptions to otherwise generally favorable criticism. For example, one respondent wrote that "the old form needs improvement, but this isn't it."

As a result of the Pay Grades 5-6 analysis, the form selected for operational implementation was Form 5. (Although Form 2a would have been similarly effective, Form 2a resembles the new operational form for evaluating Pay Grades 7-9. Since the results of previous studies suggested that the use of the same form to evaluate a wide range of pay grades contributes to inflated marking practices for ratees in the highest pay grades, there has been an effort to implement forms, for the evaluation of three different Pay Grades--7-9, 5-6 and 1-4, which are manifestly different in appearance.) Also, as part of the implementation actions, all important individual personnel selection criteria (e.g., qualifications for Reenlistment, Advancement, Honorable Discharge, etc.) were redefined in terms of a new set of alphabetic graphic type codes.

CONCLUSIONS

Graphic type scales, with general descriptors which can be employed across a wide variety of occupational specialties, can be improved for resisting rater leniency by incorporating the following design characteristics:

1. Location in the middle of the marking scale of descriptors which emphasize the high quality of the typical ratee's performance.

2. Display of a "typical" T-zone in the middle of the marking scale, and

3. Use of codes (for transcription of marks and for criteria for individual personnel actions) which connote the quality of the descriptors, as distinguished from the use of a stereotyping numerical scale.

Graphic designs can be employed which resist rater leniency, especially with the pile-up of marks in the top columns of the scale, without resorting to requirements of forced distributions (i.e., prohibiting no greater than a specified proportion of marks in each column).

REFERENCES

Bass, B, M. Reducing Leniency in Ratings. *Personnel Psychology*, 1956, *9*, 359-369.

Gordon, M. E. An Examination of the Relationship Between the Accuracy and Favorability of Ratings. *Journal of Applied Psychology*, 1972, *56*, 49-53

Robertson, D. W. *Design and Fleet Trial of an Automated Performance Evaluation Form for Chief Petty Officers.* San Diego: U. S. Naval Personnel Research Activity, May 1969. Research Report SRR 69-25.

Robertson, D. W. *Source Documents for the Automated Enlisted Performance Evaluation System.* San Diego: Naval Personnel and Training Research Laboratory, March 1972. Research Memorandum SRM 72-10.

Robertson, D. W., Royle, Marjorie H., & James, J. *Design and Fleet Trial of Automated Performance Evaluation Forms for Two Pay Grade Groups: E5-E6 and E1-E4.* San Diego: Naval Personnel and Training Research Laboratory, November 1972, Research Report SRR 73-11. (a)

Robertson, D. W., James, J., & Royle, Marjorie H. *Feasibility of Computer-Generated Data Displays in the Automated Performance Evaluation System.* San Diego: Naval Personnel and Training Research Laboratory, April 1972. Research Report SRR 72-20. (b)

Sharon, A. T., & Bartlett, C. J. Effects of Instructional Conditions in Producing Leniency on Two Types of Rating Scales. *Personnel Psychology*, 1969, *22*, 251-263.

Smith, P. C., & Kendall, L. M. Retranslation of Expectations: An Approach to the Construction of Unambigous Anchors for Rating Scales. *Journal of Applied Psychology*, 1963, *47*, 149-155.

Thomson, H. A. Comparison of Predictor and Criterion Judgments of Managerial Performance Using the Multitrait-Multimethod Approach. *Journal of Applied Psychology*, 1970, *54*, 496-502.

TABLE 1

Equivalent Numerical Scores for Comparison Of Evaluation Forms and Codes

Form	Scale Length	Code										
		Type	Scale									
1a	9	Numeric (old)	4.0	3.8	3.6	3.4	3.2	3.0	—[b]	2.6	2.0	1.0
		Numeric (new)	9	8	7	6	5	4	—	2	1	0
		Graphic			(marks only)[a]							
2a	10	Numeric (old)	4.0	3.8	3.6	3.4	3.2	3.0	2.8	2.6	2.0	1.0
		Numeric (new)	9	8	7	6	5	4	3	2	1	0
		Graphic	T05	T15	T25	T35	T50	B50	B25	B15	B5M	B5S
2b	10	Graphic	T05	T15	T25	T35	T50	B50	B25	B15	B5M	B5S
3	9	Numeric (old)	4.0	3.8	3.6	3.4	3.2	3.0	—	2.6	2.0	1.0
		Numeric (new)	9	8	7	6	5	4	—	2	1	0
		Graphic	T05	T15	T25	M5L	M5C	M5R	B25	—	B5M	B5S
5	10	Numeric (old)	4.0	3.8	3.6	3.4	3.2	3.0	2.8	2.6	2.0	1.0
		Numeric (new)	9	8	7	6	5	4	3	2	1	0
		Graphic			(marks only)[a]							
Column Scores used for analysis			9	8	7	6	5	4	3	2	1	0

[a] These subgroups were instructed to submit marks only rather than codes. The marks were coded during data processing.

[b] A dash (—) indicates an absence of that column on a particular form.

TABLE 2

Effects of Typical T-Zone

Pay Grade		Scale			
		With T		Without T	Operational (old)
7	N	762		344	2444
	$\bar{X}$	3.44		3.70	3.88
	SD	.40		.40	.21
	P<		.001		
5-6	N	165		159	
	$\bar{X}$	6.49		6.95	
	SD	1.89		1.59	—
	P<		.02		

TABLE 3

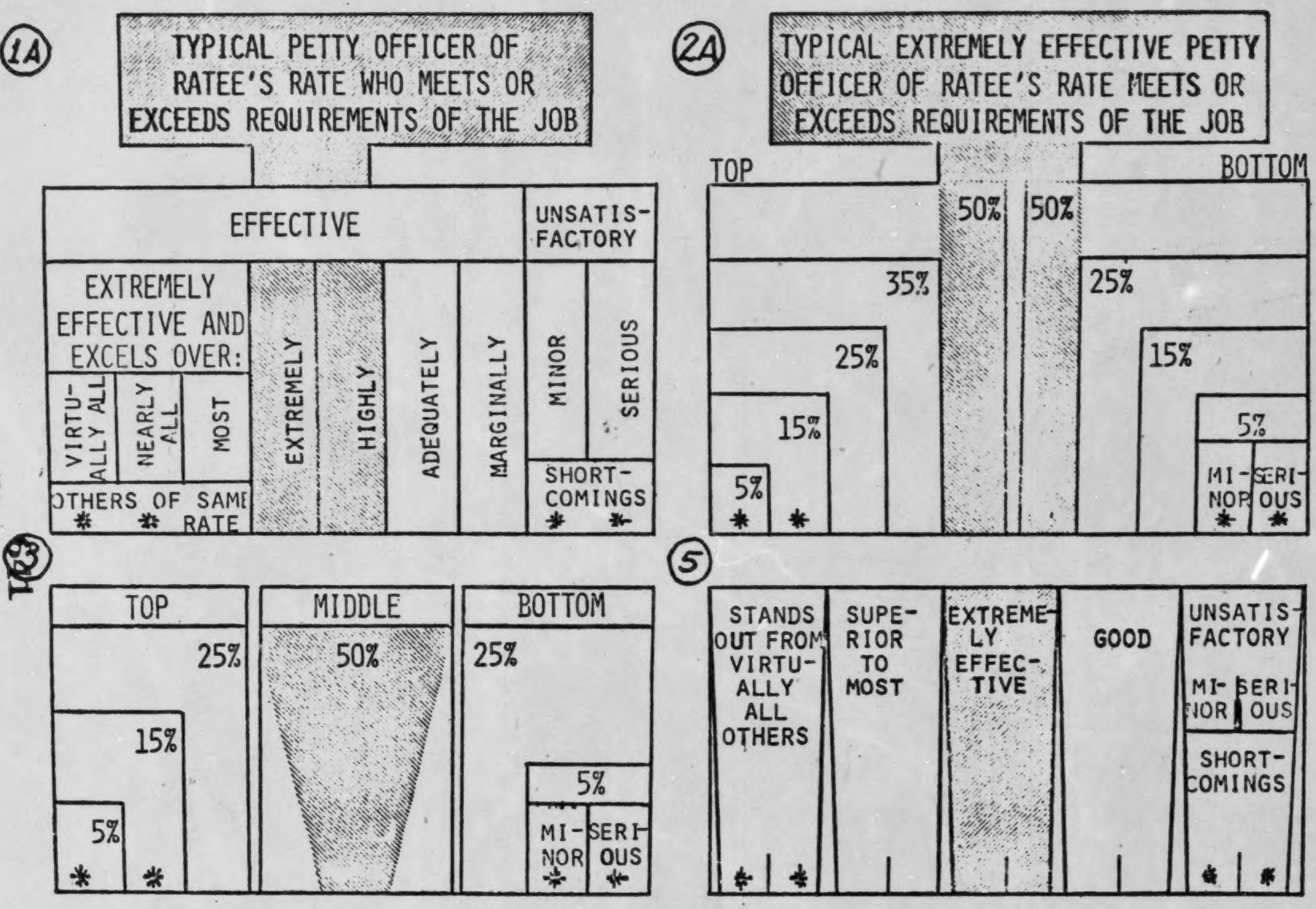
Summary of Multivariate Analyses of Variance for Individual Productivity, Pay Grades 5 and 6, Analyzed by Form Type, Coding Scheme and Rating

Source of Variation	Sum of Squares	d.f.	Mean Square	*F*	*p* less than
A. Pay Grade 5					
Within Cells	2454.203	945	2.597		
Coding Scheme	21.989	2	10.995	4.234	0.015
Numerical - present (1.0-4.0)					
Numerical - new (0-9)					
Graphic					
Form Type (1a, 2a, 3, 5)	31.630	3	10.543	4.060	0.007
Rating (AD, DC, ET, HM, PN, RM, SK)	29.165	6	4.861	1.872	0.083
Coding X Form	54.670	6	9.112	3.508	0.002
Coding X Rating	72.560	12	6.047	2.328	0.006
Form X Rating	101.794	18	5.655	2.178	0.003
Coding X Form X Rating	115.428	36	3.206	1.235	0.164
B. Pay Grade 6					
Within Cells	2968.839	1045	2.841		
Coding Scheme	45.091	2	22.546	7.936	0.001
Numerical - present (1.0-4.0)					
Numerical - new (0-9)					
Graphic					
Form Type (1a, 2a, 3, 5)	27.201	3	9.067	3.191	0.023
Rating (AD, DC, ET, HM, PN, RM, SK)	79.185	6	13.198	4.645	0.001
Coding X Form	26.735	6	4.456	1.568	0.153
Coding X Rating	38.977	12	3.248	1.143	0.320
Form X Rating	71.301	18	3.961	1.394	0.125
Coding X Form X Rating	120.361	36	3.343	1.177	0.221

TABLE 4

Rank-Order of Proportion of Individual Productivity Marks
In the Top 3 Columns for Pay Grades 5 and 6

Pay Grade 5			Pay Grade 6			Pay Grades 5 & 6 Combined		
Σ of % Top 3 Col.	Form	Coding	Σ of % Top 3 Col.	Form	Coding	Σ of % Top 3 Col.	Form	Coding
44.28	1a	Graphic	50.58	2a	Graphic	51.28	2a	Graphic
46.43	5	Graphic	54.18	5	Graphic	51.89	5	Graphic
52.11	2a	Graphic	57.95	2a	0-9	52.95	1a	Graphic
52.47	2a	0-9	60.75	5	0-9	55.27	2a	0-9
58.11	5	0-9	60.82	1a	Graphic	58.97	5	0-9
58.82	3	Graphic	66.10	1a	0-9	63.57	1a	0-9
60.38	3	1.0-4.0	72.39	1a	1.0-4.0	64.13	3	1.0-4.0
60.72	5	1.0-4.0	72.58	3	Graphic	66.94	3	Graphic
61.11	1a	1.0-4.0	72.97	3	1.0-4.0	67.55	1a	1.0-4.0
62.59	1a	0-9	75.30	3	0-9	69.92	5	1.0-4.0
63.63	2a	1.0-4.0	77.27	5	1.0-4.0	70.37	2a	1.0-4.0
85.48	3	0-9	77.77	2a	1.0-4.0	79.86	3	0-9



1A
TYPICAL PETTY OFFICER OF RATEE'S RATE WHO MEETS OR EXCEEDS REQUIREMENTS OF THE JOB
EFFECTIVE
UNSATISFACTORY
EXTREMELY EFFECTIVE AND EXCELS OVER:
VIRTUALLY ALL
NEARLY ALL
MOST
OTHERS OF SAME RATE
EXTREMELY
HIGHLY
ADEQUATELY
MARGINALLY
MINOR
SERIOUS
SHORTCOMINGS
2A
TYPICAL EXTREMELY EFFECTIVE PETTY OFFICER OF RATEE'S RATE MEETS OR EXCEEDS REQUIREMENTS OF THE JOB
TOP
BOTTOM
50%
50%
35%
25%
15%
5%
25%
15%
5%
MINOR
SERIOUS
3
TOP
MIDDLE
BOTTOM
25%
15%
5%
50%
25%
5%
MINOR
SERIOUS
5
STANDS OUT FROM VIRTUALLY ALL OTHERS
SUPERIOR TO MOST
EXTREMELY EFFECTIVE
GOOD
UNSATISFACTORY
MINOR
SERIOUS
SHORTCOMINGS

Fig. 1. Alternative evaluation scales 1a, 2a, 3 and 5 utilized in the fleet trial of automated performance evaluation forms for Pay Grades 5 and 6.

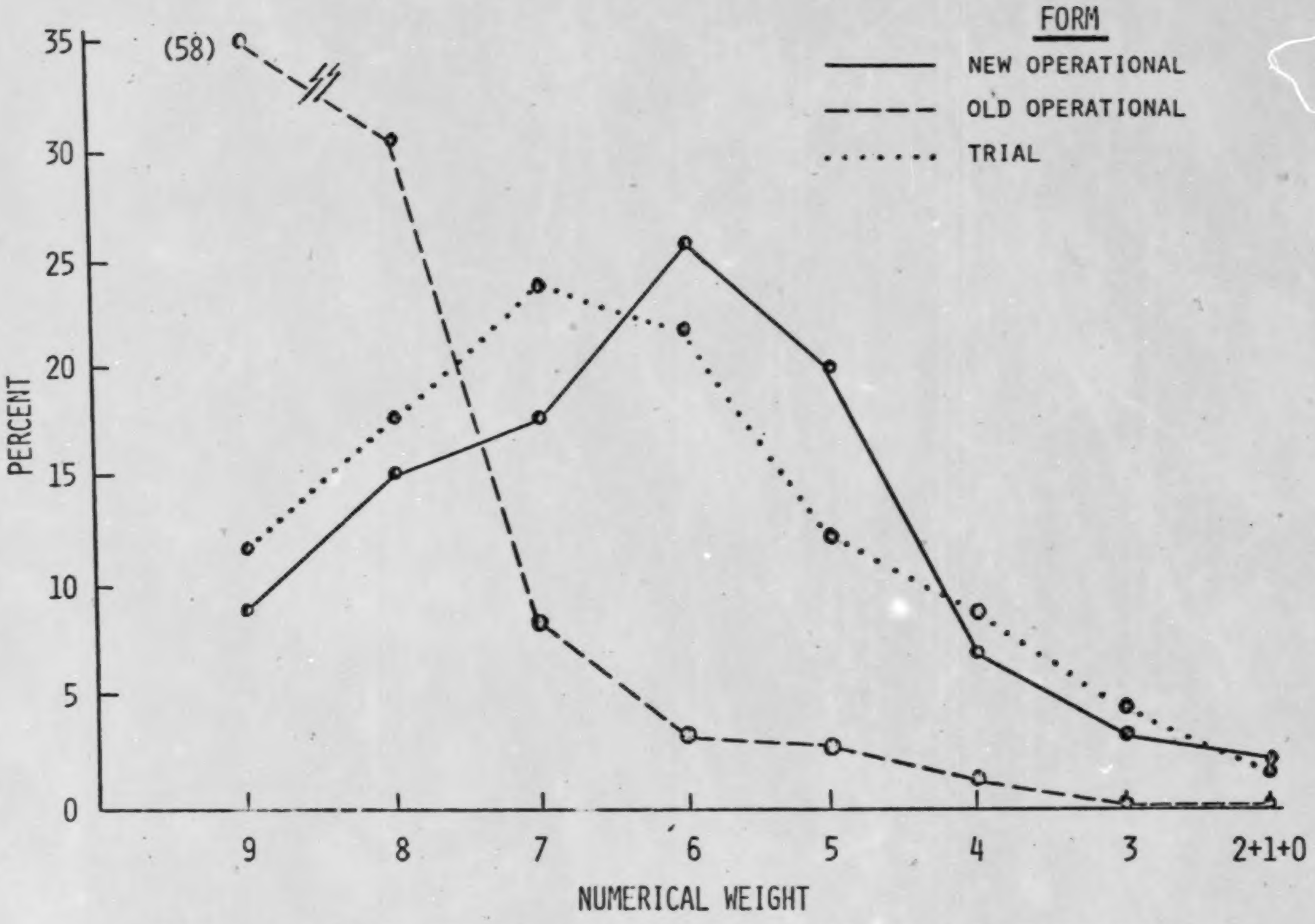


FIG. 2. COMPARISON OF MARKS ON TRIAL AND OPERATIONAL FORMS (PERFORMANCE OF DUTY ITEM, PAY GRADE 7).

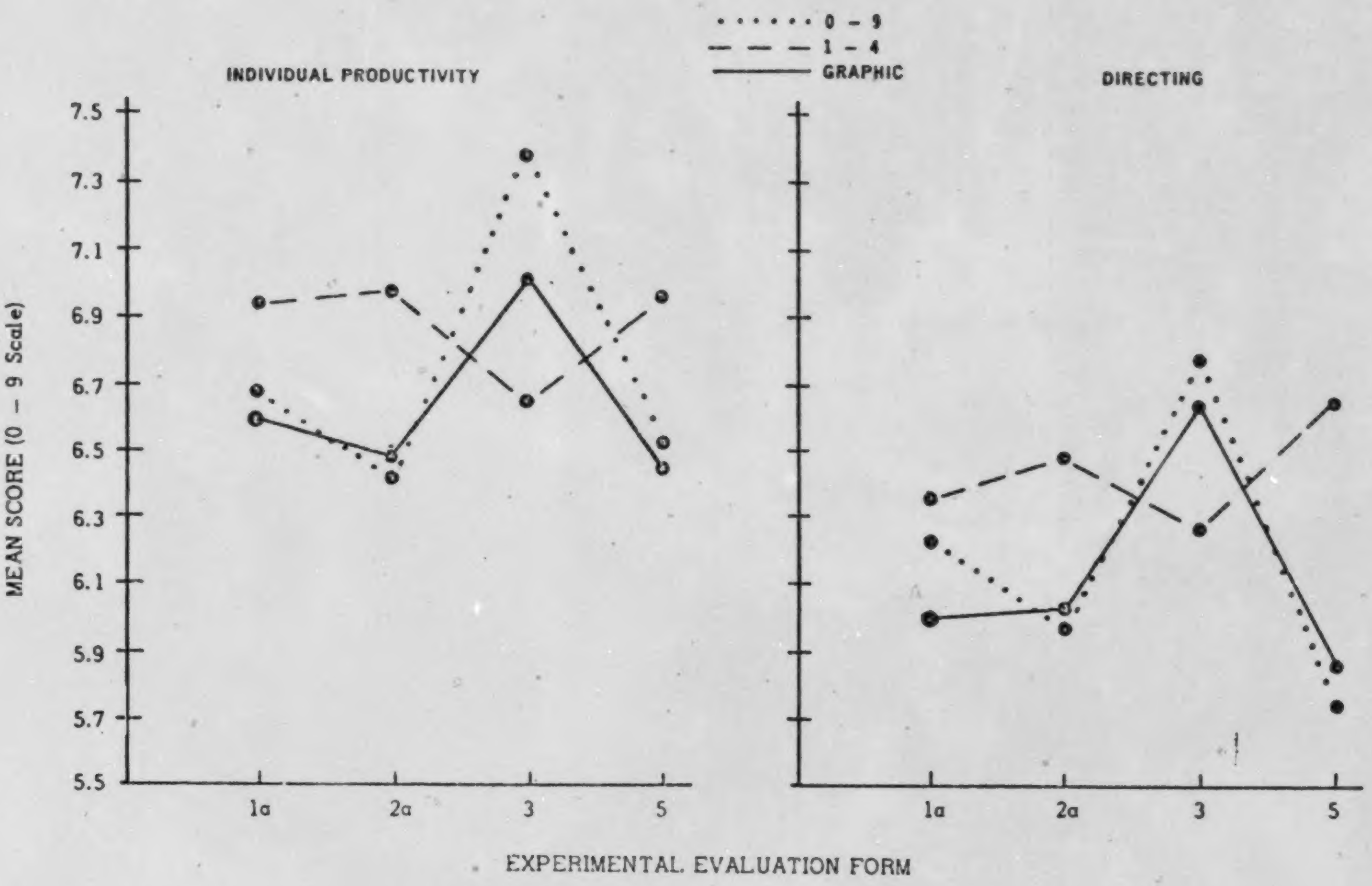


Fig. 3. Average Individual Productivity and Directing evaluation marks by form and coding type, Pay Grades 5 and 6 combined.

MINUTES

1974 MTA Steering Committee Meeting

The 1974 Military Testing Association (MTA) Steering Committee Meeting convened at the Hilton Inn West, Oklahoma City, Oklahoma on 21 October. The following members were present.

U.S. Coast Guard Institute (Coordinating Agency)

Cdr. K. M. Roughgarden, President
Mr. J. A. Burt, Secretary

U.S. Air Force Human Resources Laboratory

Col. Theodore B. Aldrich
Col. F. L. McLanathan (Retired)

U.S. Naval Education & Training Program Development Center

Capt. D. L. Hughes
Cdr. C. J. Macaluso
Mr. Walter Birdsall

U.S. Army Enlisted Evaluation Center

Dr. Clay V. Brittain

U.S. Marine Corps Headquarters

Col. J. R. Gallman, Jr.

Royal Australian Air Force

Sq. Ldr. W. J. Watson

Canadian Armed Forces

Lt. Col. G. J. Carpenter

U.S. Armed Forces Vocational Testing Group

Col. R. S. Hogatt, USAF

U.S. Naval Personnel Research and Development Center

Mr. P. P. Foley

1. The first order of business was the election of a Chairman of the Steering Committee, in accordance with the By-Laws adopted last year in San Antonio. After a brief discussion, Col. Aldrich moved that Article V-C of the By-Laws be amended to read that: "The Chairman of the Steering Committee shall be appointed by the President of the Association."

2. The results of the Air Force survey of the Steering Committee on changing the name of the Association were reported. The nine names on the ballot produced no clear consensus. In fact, five write-in names were received. The ensuing discussion culminated in a motion by Lt. Col. Carpenter that the matter of a name change be tabled. The motion was seconded and passed unanimously.

3. Application for Steering Committee memberships were recognized from Col. Hogatt, USAF, representing the Armed Forces Vocational Testing Group, and by Mr. P. Foley, representing the Naval Personnel Research and Development Center.

a. Cdr. Macaluso, USN, made a motion that the Armed Forces Vocational Testing Group be accepted as a full member of the Steering Committee. The motion was seconded and carried unanimously.

b. Col. Hogatt made a motion that the Naval Personnel Research and Development Center be made a full member of the MTA Steering Committee. The motion was seconded and carried unanimously.

4. There was a brief discussion on Article VII-B of the By-Laws on the rotational schedule of coordinating the annual conference. Consensus was reached that the rotating schedule should be retained in its current form.

5. As the Army's representative and next year's MTA Host, Dr. Clay Brittain announced the conference will be held at Indianapolis; however, the specific place and date have yet to be determined.

6. No formal nominations were received for the Harry H. Greer Award. Col. McLanathan recommended the Steering Committee give serious consideration to a nomination for the award next year.

7. The meeting adjourned at 1730.

8. The Steering Committee reconvened at 0900, Thursday, 24 October, to consider the possibility of the formation of a separate division of the Association for occupational analysis. The sense of the Committee was that the purpose of a separate division could be better served by more and better pre-convention publicity. It was suggested to next year's conference coordinator, the U.S. Army Enlisted Evaluation Center, that a separate day be set aside on the program for occupational analysis papers. The Committee readjourned at 0945.

Minutes of the 1974 MTA Business Meeting

The meeting was called to order at 1515, 24 October, by the MTA secretary, Mr. John Burt.

1. The minutes of the Steering Committee were presented to the general assembly. On motion of the Steering Committee and by unanimous consent of assembly, Article V-C of the By-Laws was amended to read that "The Chairman of the Steering Committee shall be appointed by the President of the Association."

2. A floor motion that a survey of desired paper topics be distributed with the conference proceedings carried unanimously. The surveys will be returned to next year's conference coordinator, the U.S. Army Enlisted Evaluation Center.

3. Recommendations were solicited from the floor concerning the conduct and format of next year's conference. The following were proposed:

    a. One hour forums will be held to discuss the problems of the "operators" in the field of personnel assessment.

    b. Availability of a panel of experts in particular areas to respond to questions from the audience. Such areas could include occupational analysis, test construction, etc.

    c. The type size of view graphs be enlarged.

The meeting adjourned at 1540.

BY-LAWS OF THE MILITARY TESTING ASSOCIATION

Article I - Name

The name of this organization shall be the Military Testing Association.

Article II - Purpose

The purpose of this Association shall be to:

A. Assemble representatives of the various armed services of the United States and such other nations as might request to discuss and exchange ideas concerning assessment of military personnel.

B. Review, study, and discuss the mission, organization, operations, and research activities of the various associated organizations engaged in military personnel assessment.

C. Foster improved personnel assessment through exploration and presentation of new techniques and procedures for behavioral measurement, occupational analysis, manpower analysis, simulation models, training programs, selection methodology, survey and feedback systems.

D. Promote cooperation in the exchange of assessment procedures, techniques and instruments.

E. Stimulate and coordinate research in the various areas of military personnel systems.

F. Promote the assessment of military personnel as a scientific adjunct to modern military personnel management within the military and professional communities.

Article III - Participation

The following categories shall constitute membership within the MTA:

A. Primary Membership.

1. All active duty military and civilian personnel permanently assigned to an agency of the associated armed services having primary responsibility for assessment for personnel systems.

2. All civilian and active duty military personnel permanently assigned to an organization exercising direct command over an agency of the associated armed services holding primary responsibility for assessment of military personnel.

B. Associate Membership.

1. Membership in this category will be extended to permanent personnel of various governmental, educational, business, industrial and private organizations engaged in activities that parallel those of the primary membership. Associate members shall be entitled to all privileges of primary members with the exception of membership on the Steering Committee. This restriction may be waived by the majority vote of the Steering Committee.

Article IV - Dues

No annual dues shall be levied against the participants.

Article V - Steering Committee

A. The governing body of the Association shall be the Steering Committee. The Steering Committee shall consist of voting and non-voting members. Voting members are primary members of the Steering Committee. Primary membership shall include:

1. The Commanding Officers of the respective agencies of the armed services exercising responsibility for personnel assessment programs.

2. The ranking civilian professional employees of the respective agencies of the armed service exercising primary responsibility for the conduct of personnel assessment systems. Each agency shall have no more than two (2) professional civilian representatives.

B. Associate membership of the Steering Committee shall be extended by majority vote of the committee to representatives of various governmental, educational, business, industrial and private organizations whose purposes parallel those of the Association.

C. The Chairman of the Steering Committee shall be appointed by the President of the Association.

D. The Steering Committee shall have general supervision over the affairs of the Association and shall have the responsibility for all activities of the Association. The Steering Committee shall conduct the business of the Association in the interim between annual conferences of the Association by such means of communication as deemed appropriate by the President or Chairman.

E. Meeting of the Steering Committee shall be held during the annual conferences of the Association and at such times as requested by the President of the Association or the Chairman of the Steering Committee. A majority of the members of the Steering Committee shall constitute a quorum.

Article VI - Officers

A. The officers of the Association shall consist of a President, Chairman of the Steering Committee and a Secretary.

B. The President of the Association shall be the Commanding Officer of the armed services agency coordinating the annual conference of the Association. The term of the President shall begin at the close of the annual conference of the Association and shall expire at the close of the next annual conference.

C. It shall be the duty of the President to organize and coordinate the annual conference of the Association held during his term of office, and to perform the customary duties of a president.

D. The Secretary of the Association shall be filled through appointment by the President of the Association. The term of office of the Secretary shall be the same as that of the President.

E. It shall be the duty of the Secretary of the Association to keep the records of the Association, and the Steering Committee, and to conduct official correspondence of the Association, and to issue notices for conferences. The Secretary shall also perform such additional duties and take such additional responsibilities as the President may delegate to him.

Article VII - Meetings

A. The Association shall hold a conference annually.

B. The annual conference of the Association shall be coordinated by the agencies of the associated armed services exercising primary responsibility for military personnel assessment in order of the following rotating schedule:

United States Army
United States Marine Corps
United States Navy
United States Air Force
United States Coast Guard

C. The annual conference of the Association shall be held at a time and pace determined by the coordinating agency. The membership of the association shall be informed at the annual conference of the place at which the following annual conference will be held. The coordinating agency shall inform the Steering Committee of the time of the annual conference not less than six (6) months prior to the conference.

D. The coordinating agency shall exercise planning and supervision over the program of the annual conference. Final selection of program content shall be the responsibility of the coordinating organization.

E. Any other organization desiring to coordinate the conference may submit a formal request to the Chairman of the Steering Committee, no later than 18 months prior to the date they wish to serve as host.

Article VIII - Committees

A. Standing committees may be named from time to time, as required, by vote of the Steering Committee. The chairman of each standing committee shall be appointed by the Chairman of the Steering Committee. Members of standing committees shall be appointed by the Chairman of the Steering Committee in consultation with the chairman of the committee in question. Chairmen and committee members shall serve in their appointed capacities at the discretion of the Chairman of the Steering Committee. The Chairman of the Steering Committee shall be an ex officio member of all standing committees.

B. The President with the counsel and approval of the Steering Committee may appoint such ad hoc committees as are needed from time to time. An ad hoc committee shall serve until its assigned task is completed or for the length of time specified by the President in consultation with the Steering Committee.

C. All standing committees shall clear their general plans of action and new policies through the Steering Committee, and no committee or committee chairman shall enter into relationships or activities with persons or groups outside of the Association that extend beyond the approved general plan of work without the specific authorization of the Steering Committee.

D. In the interest of continuity, if any officer or member has any duty elected or appointed placed on him, and is unable to perform the designated duty, he should decline and notify at once the officers of the Association that he cannot accept or continue said duty.

Article IX - Amendments

A. Amendments of these By-Laws may be made at any annual conference of the Association.

B. Amendments of the By-Laws may be made by majority vote of the assembled membership of the Association provided that the proposed amendments shall have been approved by a majority vote of the Steering Committee.

C. Proposed amendments not approved by a majority vote of the Steering Committee shall require a two-thirds vote of the assembled membership of the Association.

Article X - Voting

All members in attendance shall be voting members.

Article XI - Enactment

These By-Laws shall be in force immediately upon acceptance by a majority of the assembed membership of the Association and/or amended (in force 2 November 1973).

LIST OF CONFEREES

ALDRICH, COL Theodore B.
AFHRL, Lackland AFB TX 78148
AMOS, CWO3 E. L.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
ANDERSON, COL Charles C., Jr. (Ret.)
Educational Testing Service, 960 Grove St., Evanston IL 60201
ANSBRO, Mr. Thomas M.
Service School Command, Naval Training Center, Orlando FL
ARNOLD, Mr. George C.
Armory School, Ft. Knox KY
ASA-DORIAN, Mr. Paul V.
Fleet Anti-Sub Warfare Training Center, Code 32, San Diego CA 92147
BEAVERS, LCDR E. L.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
BEGLAND, CAPT Robert R.
Faculty Development Division, USAIS, Ft. Benning GA 31905
BERGER, Mr. B. Michael
MILPERCEN, PMDD, MOR Division, 2461 Eisenhower, Alexandria VA 22331
BIRDSALL, Mr. Walter W.
Naval Education & Training Program Development Center, Bldg 923, Rm 111, Ellyson, Pensacola FL 32509
BLAND, LCDR R. D.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
BOBROWSKI, CWO3 T. J.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
BOHN, Ms. Carole A.
Texas Technical University, Psychology Dept., Lubbock TX 79409
BONKER, RMCM D. L.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
BOYD, Dr. Joseph L.
Educational Testing Service, Princeton NJ 08540
BRADBURY, SSCM H. E.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
BRASHEAR, Mr. Fenton W.
US Army Ordnance Center & School, Office of Training and Education, Aberdeen Proving Ground MD 21005
BRAUN, Mr. Frank B.
Data-Design Laboratories, 15 Koger Executive Center, Suite 140, Norfolk VA 23502
BRITTAIN, Mr. Clay V.
US Army Enlisted Evaluation Center, Ft. Benjamin Harrison IN 46249
BROGA, LT D. C. III
U.S. Coast Guard, COMDT (G-PE-1), 400 Seventh St., SW, Washingon DC 20590
BROWN, CDR Bruce W.
Chief Naval Education & Training, Code N-34, NAS Pensacola FL 32508
BUONOCORE, Mr. George V.
US Postal Service, 475 L'enfant Plaza West, SW, Rm 7210, Wash DC 20260
BURNS, Ms. Darla J.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169

BURT, Mr. John A.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
BUSBY, CDR C. W.
USCG Training Center, Petaluma CA 94952
BUSH, CWO3 Billy J.
USCG Training Center, Petaluma CA 94952
CALLAHAN, Mr. M. D.
Technical Director, NOTAP, Bldg 1304 BAFP, Washington DC 20332
CARDACI, LT Phillip J.
USCG, COMDT (G-PTE/72), 400 Seventh St., SW, Washington DC 20590
CARLTON, CDR J. C.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
CARPENTER, LT COL G. J.
Royal Military College of Canada, Kingston, Ontario, CANADA K7L2W3
CASH, CMSgt Robert L.
USAF Senior NCO Academy, 2750 Argyle Rd., Montgomery AL 36111
CAVINESS, Dr. James A.
ARI, 2101 Coventry Dr., Columbus GA 31904
CHASE, LTJG P. K.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
CHENEY, LT Knight D.
USCG, COMDT (G-RT/81), 400 Seventh St., SW, Washington DC 20590
CONNER, DTCS J. E.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
CORBEIL, HMCM S. F.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
CORMACK, CDR B.
Director Occupational Analysis, National Defence HQ, Ottawa, Ontario, CANADA K1A0K2
COWAN, Mr. Joseph
USCG, COMDT (G-P-1/62), 400 Seventh St., SW, Washington DC 20590
CROSS, PACM D. E.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
CROSS, LCDR N. J.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
DALIMOT, SSCS P. C.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
DANN, Ms. Joyce E.
NPRDC, San Diego CA 92152
DASH, CDR R. E.
USCG Training Center, Cape May NJ 08204
DEBERG, CAPT Oak H.
Aeronautical Systems Division, ASK/ENCTH, Wright-Patterson AFB OH 45433
DEES, Dr. James W.
HumRRO, P.O. Box 428, Ft. Rucker AL 36360
DUBOSKY, Mr. Stephen W.
Department of Resident Training Management, US Army Aviation Center, Ft. Rucker AL 36360
DUGAN, ASMCM C. J.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169

DURFEY, RADM Robert W.
US Coast Guard Headquarters, Washington DC 20590
EHRHARDT, Mr. Louis E.
Naval Millile Center, Code N342, Pt. Mugu CA 93042
ELEY, ASM1 R. J.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
ELLARD, LTJG R. B.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
ELSTER, Dr. D.
Naval Postgraduate School, Monterey CA
EYSMANS, ENS T. M. S.
USCG Training Center, Cape May NJ 08204
FISCHL, Dr. M. A.
US Army Research Institute, 1300 Wilson Blvd., Arlington VA 22209
FLANAGAN, LTJG T. J.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
FOLEY, Mr. Paul P.
Naval Personnel Research Development Center, San Diego CA 92152
FORD, Mr. J. Patrick
HumRRO Div. 2, Ft. Knox KY 40121
FRICKE, LT COL Jurgen
Federal Armed Forces, West Germany
GALLMAN, COL J. R.
HQ, US Marine Corps, Office of Manpower Utilization, Quantica VA 22134
GERALDI, Mr. Vincent A.
US Army Communications Electronics School, Ft. Monmouth NJ 07703
GERBER, Mr. J. E. Jr.
US Army Infantry School, STSH-DOI-ED, Ft. Benning GA 31905
GIERHART, Mr. H. S.
USCG, COMDT (G-P-1/62), 400 Seventh St., SW, Washington DC 20590
GILBERT, Dr. Arthur C. F.
US Army Research Institute, 1300 Wilson Blvd., Arlington VA 22209
GLADWIN, MKCM R. E.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
GMEINDER, Dr. Susan
Naval Instructional Training Development Center, San Diego CA 92133
GODFREY, Mr. Richard
Community College, B661, Randolph AFB TX 78148
GOSS, CDR James E.
Naval Education & Training Program Development Center, Pensacola FL
GRADY, LCDR F. J. III
USCG Institute, P. O. Substation 18, Oklahoma City OK 73169
GREGOIRE, LCDR Harv
Texas Technical University, Psychology Department, Lubbock TX 79409
GRINDSTAFF, LT Terry L.
USCG, COMDT (G-BBE), 400 Seventh St., SW, Washington DC 20590
GULDENSCHUH, LT C. J.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
HAHN, Mr. C. P.
American Institutes for Research, 3301 New Mexico Ave., Washington DC
HALDEMAN, CWO3 J. A. III
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169

HALTRECHT, Dr. Ed
Ontario Hydro, 620 University Ave., Rm GU146, Toronto, Ontario, CANADA
HARRIS, Mr. James H.
HumRRO Div 2, Ft. Knox KY 40121
HARRIS, ADCM J. R., Jr.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
HARRIS, Mr. William S.
National Test Center (USPS), 300 N. Los Angeles, Los Angeles CA 90012
HARTSELL, RDCS G. K.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
HAZARD, GMC R. K.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
HECKINGER, Ms. Amy R.
US Intelligence School & Center, Evaluation Branch, Ft. Huachuca AZ
HINSON, CDR G. C.
USCG Training Center, Governors Island NY 10004
HOGGATT, COL Ralph S.
Armed Forces Vocational Testing Group, Randolph AFB TX 78148
HOLT, Dr. Howard O.
57 Hillcrest, Summit NJ 07901
HORAN, LTJG W. T.
USCG, COMDT (G-P-1/62), 400 Seventh St., SW, Washington DC 20590
HOYT, Dr. William G.
System Development Corporation, 2500 Colorado Ave., Santa Monica CA 90406
HUBER, ETC T. R.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
HUGHES, CAPT D. L.
Naval Education & Training Program Development Center, Pensacola FL 32509
HYATT, SQ LDR Christopher J. (RAF)
Aeronautical Systems Division, ASD/ENCTH, Wright-Patterson AFB OH 45433
IMMER, LT H. L.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
IRISH, Ms. Sally C.
USCG, COMDT (G-BBE), 400 Seventh St., SW, Washington DC 20590
ISRAEL, Ms. Deborah
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
JACOBS, Mr. Robert S.
Aviation Research Laboratory, Savoy IL
JOHANSSON, Dr. Charles B.
4401 W. 76 Street, Minneapolis MN 55435
JOHNSON, ETCM B. H., Jr.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
JOHNSON, BMCS R. P.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
JOHNSON, Mr. Eugene
Army Research Institute, 1300 Wilson Blvd., Arlington VA 22209
JOHNSON, Mr. Kirk
Naval Personnel Research & Development Center, Memphis TN
JOHNSON, Mr. Robert N.
US Army Institute of Administration, Ft. Benjamin Harrison IN 46216

JONES, LTJG J. D.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
KELLEHER, Ms. Anna May
USCG Training Center, Governors Island NY 10004
KERR, Dr. Norman J.
NAS Memphis (75), Millington TN 38054
LANTERMAN, Mr. Richard S.
USCG, COMDT (G-P), 400 Seventh St., SW, Washington DC 20590
LEE, AMC R. E.
USCG Institute, P.O. Substation 18, Washington DC 20590
LEWIS, Ms. Nancy A.
Occupational Measurement Center, OM, Lackland AFB TX 78236
LEWISKI, Dr. Raymond
Extension Course Institute, EDXV, Gunter AFS AL
LINDSAY, LT Alan W.
CNET Staff, Code N-3343, NAS Pensacola FL 32508
MACALUSO, CDR C. J.
Naval Education and Training Program Development Center, Pensacola FL
MAIER, Dr. Milton H.
Army Research Institute, 1300 Wilson Blvd., Arlington VA 22209
MANN, CWO3 B. E.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
MCBRIDE, Mr. Curtis D.
US Army Field Artillery School, Evaluation Branch, Ft. Sill OK 73603
MCCLELLAN, Dr. Willard C.
Naval Education & Training Program Development Center, Pensacola FL
MCCULLAH, LCDR R. D.
Behavioral Science Dept., USN Academy, Annapolis MD 21402
MCDEVITT, LT D. E.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
MCLANATHAN, COL Frank L.
St. Mary's University, San Antonio TX 78201
MILLER, CDR James P.
Navy Occupational Task Analysis Program, Bolling AFB, Washington DC
MILLER, FTC T. R.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
MIRABELLA, Dr. Angelo
Army Research Institute, 1300 Wilson Blvd., Arlington VA 22209
MOLZ, LTJG Richard E.
USCG, COMDT (G-RT/81), 400 Seventh St., SW, Washington DC 20590
MOONAN, Dr. William J.
Navy Program Research Development Center, San Diego CA 92152
MOORE, LTJG R. P.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
MORHARDT, Ms. Francia
USPS, 475 L'enfant Plaza West, SW, Washington DC 20260
MULLER, Mr. Richard
Naval Education and Training Program Development Center,
Pensacola FL 32508

NICHOLL, CDR Colin J.
Naval Manpower Utilization Unit, HMS VERNON, Portsmouth, ENGLAND
NODDIN, Mr. Ernest M.
Submarine Medical Research Lab, USN Submarine Base, Groton CN 06340
NOWE, Mr. C. Edward
AAC-230, FAA Center, Oklahoma City OK 73125
OETTING, LT COL George
Extension Course Institute, Gunter AFS AL 36118
OSBORN, Mr. William C.
HumRRO Div. 2, Ft. Knox KY 40121
PARKER, RMC B. D.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
PARNELL, MKCS J. C.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
PAUL, Ms. L. J.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
PEARSON, LT T. W., Jr.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
PHILLIPS, Mr. Robert H.
Reserve Training Center, Yorktown VA 23690
PORTER, Mr. G.V., Jr.
Exam Division, USAFA/RRE, USAF Academy CO 80840
POST, TTCM O. J.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
POVENMIRE, Mr. H. K.
Aviation Training Center, USCG, Mobile AL 36608
PRATT, Mr. W. H.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
PRICE, YNCM W. A.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
PRIM, MAJ David F.
US Army Military Police School, Ft. Gordon GA 30905
QUINN, LCDR J. C.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
REED, YNCS R. E.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
RIEBEL, CWO2 H. L.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
RIEDMULLER, BMCM C., Jr.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
RIGBY, CAPT Randall
University of Oklahoma, Norman OK 73069
RILEY, Mr. Eldon W.
Chief of Naval Education & Training, Code AJ-401, Pensacola FL 32508
RITCHIE, QMCM B. R.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
ROBERTS, Mr. C. D.
Army Enlisted Evaluation Center, Ft. Benjamin Harrison IN
RODGERS, STCM J. D.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
ROUGHGARDEN, CDR K.M.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169

ROYLE, Ms. Marjorie H.
Navy Personnel Research & Development Center, San Diego CA 92152
RUBERTON, Mr. Louis
Department of Army Headquarters, Washington DC
SANDS, Mr. William A.
Navy Personnel Research & Development Center, San Diego CA 92152
SARGENT, SKC Q. S.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
SCANLAND, Dr. Francis W.
Chief of Naval Education & Training, Pensacola FL 32508
SCHMIDT, CWO3 D. L.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
SHIELDS, LCDR W. S.
Royal Military College of Canada, Kingston, Ontario, CANADA K7L4X5
SIMPSON, Mr. Robert C., Jr.
USPS, 300 N. Los Angeles, Los Angeles CA 90012
SLACK, MAJ W. E.
Office of Manpower Utilization, Marine Corps Base, Quantico VA 22134
SMITH, Mr. Carl R.
USA Enlisted Evaluation Center, Ft. Benajmin Harrison IN 46249
SMITH, Mr. Mabron H.
US Army Missile & Munition Center & School, Redstone Arsenal AL
SMITH, Dr. Margaret
Naval Education & Training Program Development Center, Pensacola FL
SMITH, Mr. Ronald H.
Naval Education & Training Program Development Center, Pensacola FL
STEFL, Mr. William A.
USPS, 475 L'enfant Plaza West, Washington DC 20260
STEVENSON, MAJ Hugh T.
US Army Engineer School, Ft. Belvoir VA
STOCK, Dr. William E.
Minnesota Department of Education, St. Paul MN 55101
SULLIVAN, Mr. Denis J.
USASA Training Center, Ft. Devens MA 01433
TEASLEY, Mr. Keith
AAC-230, FAA Center, Oklahoma City OK 73125
THAYER, Mr. D. M.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
THOMAS, Dr. David B.
CAIL, University of Iowa, Iowa City IA 52242
TINDALL, Mr. J. E.
Data-Design Laboratory, Norfolk VA 23502
TOOTLE, CAPT A. Eugene
AIAOS, Box 2202, Maxwell AFB AL 36112
TUTTLE, Dr. Thomas C.
Westinghouse Behavioral/Safety Center, Box 948, Columbia MD 21044
VAUGHN, 1LT David S.
USAF Occupational Measurement Center, Lackland AFB TX 78236
WADLIN, Ms. J. L.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
WALKER, LCDR C. Lee
Central Test Site for PTEP, Dam Neck VA 23461

WARNOCK, Dr. A. Timothy
Air War College Associate Program, Maxwell AFB AL 36112
WARTHEN, CAPT Donald
CO, Service School Command, Naval Training Center, Orlando FL 32813
WATSON, SQ LDR W. J.
Royal Australian Air Force, AFHRL/OR, Lackland AFB TX 78236
WAUGH, Dr. Dorothy E.
National Security Agency, Ft. George C. Meade MD
WEIR, YNC R. D.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
WEST, Ms. Anita S.
University of Denver Research Institute, Denver CO 80210
WHEAT, Ms. P. L.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
WILLIAMS, Prof. Ralph E.
University of Denver Research Institute, Denver CO 80210
WILLIAMS, Mr. Rayburn A.
Instructional Program Development Section, Pensacola FL 32508
WILLING, Mr. R. C.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
WILMOTH, LCDR M. E.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
WINIEWICZ, Mr. Casimer S.
US Civil Service Commission, Chicago IL 60607
WOHLFARTH, LT W. G.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
WOOD, Dr. Adele H.
Naval Health Science Education & Training Command, National Naval Medical Center, Bethesda MD
WRIGHT, LT Richard W.
USCG, COMDT (G-PE), 400 Seventh Street, SW, Washington DC 20590
WYLIE, Dr. Peter B.
HumRRO Div 2, Ft. Knox KY 40121
YELLEN, Mr. Ted
Navy Personnel Research & Development Center, San Diego CA 92152
YELTON, YNC E. F. Jr.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
ZABLE, CWO3 G. E.
USCG Institute, P.O. Substation 18, Oklahoma City OK 73169
ZECHLIN, Mr. Frank F.
Data-Design Laboratory, 7925 Center Ave., Cucamonga CA
ZURO, COL Joseph E.
Air War College (USAF), Maxwell AFB AL 36112

INDEX

PAGE

BIRDSALL, Mr. Walter
The Perspective of a Testing Psychologist Who Conducts Orientation Programs for New Test Constructors 502

BOHN, Ms. Carole
An Automated Aircrew Station Anthropometric Unit . . . 256

BOYD, Mr. Joseph L.
The Effect of Format Change on Test Item Difficulty . . 184

BRASHEAR, Mr. F. W.
Student Response in Television Instruction 280

BRAUN, Mr. F. B. and TINDALL, Mr. J. E., Jr.
A New Sequential Multiple-Choice Skill-Related Testing Device 194

CALLAHAN, Mr. M. D.
Naval Occupational Task Analysis Program 346

CARPENTER, LT COL G. J.
The Development of an Instrument for Student Evaluation of Course and Instructor 540

CAVINESS, Dr. J. A.
Assessment and Counseling for Leadership Development . 442

CORMACK, CDR B.
One of the Missing Links 27

DANN, Dr. Joyce E.
Clemans' Lambda-Based Scales of the Navy Vocational Interest Inventory 100

DEES, Dr. James W.
Structural Task Analysis--The Bridge Between Selection and Training 31

DOWNEY, Dr. Ronald G.
Factor Analysis of Leadership Measures From a Training Program 452

DRISKILL, Dr. Walter E.
Occupational Analysis in the United States Air Force . 340

DURFEY, RADM R. W.
Keynote Address 12

PAGE

EHRHARDT, Mr. Louis E.
Development of an RPV Operator Trainer 149

ELSTER, Dr. R. S. and GITHENS, W. H.
Selection of Officer/Students for Graduate Education 235

FISCHL, Dr. M. A.
Interest Measurement for Army Classification 127

FRICKE, LT COL Jurgen
Vocational Advancement in the Federal Armed Forces . . . 335

GILBERT, Dr. Arthur C. F.
Classifications of Army Officer Duty Positions: Implications for Training 559

HAHN, Mr. Clifford P.
Development of Criteria Ratings to Measure Job Performance 596

HARRIS, ADCM J. R.
The Perspective of a Non-Commissioned Officer Who Is Actually Constructing Tests 514

HELMICK, CAPT John C., Jr.
The Common Bond in Peer and Preceptor Ratings 314

HERMANSEN, Mr. A. G. and BRITTAIN, Dr. Clay V.
A 1" x 3½" Keyhole on the World of MOS Testing 187

HOYT, Mr. W. G. and BAKER, Mr. James D.
The Use of Tactical Computers to Provide Weapons and Tactics Training to Combat NCOs: Results of a Field Test 425

HYATT, SQ LDR C. J. and DEBERG, CAPT O. H.
Performance Evaluation: The Use of Scoring Systems in Adaptive Training 155

IRISH, Ms. Sally C.
Boating Safety Education: Three Routes to a Common Goal 274

JACOBS, Mr. Robert S.
Simulator Motion As a Factor in Flight Simulator Training Effectiveness 135

JOHANSSON, Dr. Charles B.
Interpretive Report for the Strong-Campbell Interest Inventory 46

PAGE

JOHNSON, Mr. Eugene and MIRABELLA, Dr. Angelo
Test Fidelity: Congruence of Test Conditions
with Actual Army Tasks. 590

JOHNSON, Mr. Robert N. and MALONE, MAJ R. F.
Test Results: The Key to the Management of Training. . 372

KAAPKE, COL Lyle D.
The Interface of Testing and Training: Air Force
Experience 1972 - 1974. 386

LEWIS, Ms. Nancy A.
Use of Occupational Survey Data in Developing
Outlines for Specialty Knowledge Tests. 178

LEWISKI, Dr. Raymond L.
Public Policy Concerning the Use of Tests 289

MOONAN, Dr. William J.
Contributions to Objective Measurement and
Evaluation of Trainee Competency. 393

NICOLL, CDR Colin J.
Occupational Analysis in the Royal Navy 357

OETTING, LT COL George D.
The Care and Feeding of Test Constructors 518

OSBORN, Mr. W. C., HARRIS, Mr. J. H., and FORD, Mr. J. P.
Functionally Integrated Performance Testing 582

POVENMIRE, Mr. H. Kingsley
Automated Performance Measurement Used in Quality
Control of the U. S. Coast Guard Aviator Training
System. 243

ROYLE, Ms. Marjorie H. and ROBERTSON, Mr. David W.
Scaling Techniques to Reduce Rater Leniency on
Navy Enlisted Performance Evaluation Forms. 616

SANDS, Mr. William A. and McCULLAH, LCDR R. D.
Development and Evaluation of Strong Vocational
Interest Blank Scales for Academic Major Areas
at the U. S. Naval Academy. 112

SCANLAND, Dr. F. W.
The Conflicting Demands of Go-No-Go Performance
Criteria and the Requirements for Rank Ordering
Students. 496

PAGE

SELLMAN, MAJ W. S. and FUGILL, SQ LDR J. W. K
Development and Application of an Algorithm for Translating Job Information into Technical Training Requirements 17

The Royal Australian Air Force Occupational Measurement Program 362

SHIELDS, LCDR W. S.
Prediction from Contingency Tables Using Joint Likelihoods 261

SLACK, MAJ W. E.
Response Bias in Worker Task Inventories. 464

SMITH, Mr. Carl R.
The Perspective of a Civilian Test Psychologist Working with Test Constructors. 510

TIMMERMAN, Dr. F. W. and RIGBY, CAPT Randall
Problem Checklists, A Pragmatic Approach for Fostering Unit Effectiveness. 552

TUTTLE, Dr. Thomas C.
Behavioral Toxicology: A New Frontier for Behavioral Testing. 212

VAUGHN, 1LT David S.
The Perspective of a Military Supervisor of Test Constructors Who Are Temporarily Brought in From the Field to Write Tests. 506

WALKER, LCDR Clarence L.
Task Analysis Based Testing as a Measure of Training Effectiveness. 381

WEST, Ms. Anita S. and WILLIAMS, Prof. Ralph E.
Entrance Scores and Training Performance for Two Groups of Coast Guard School Input. 225

WILLING, Mr. R. C. and McDEVITT, LT D. E.
U. S. Merchant Marine Licensing Examinations: The Modernization of a Traditional Mission. 576

YELLEN, Mr. Theodore
An Experimental Assessment Procedure for Evaluating Overseas Performance 459